.UME 108 PART B NUMBER 40 JULY 1961

The Proceedings

OF

THE INSTITUTION OF ELECTRICAL ENGINEERS

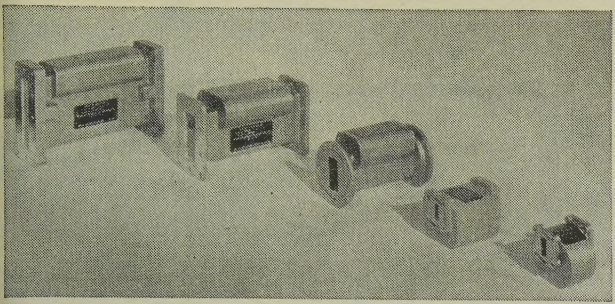
FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

ELECTRONIC AND COMMUNICATION ENGINEERING
(INCLUDING RADIO ENGINEERING)

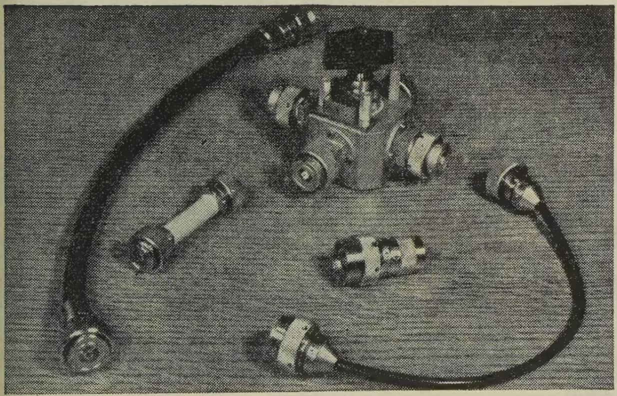
Price Ten Shillings and Sixpence

Ferrite Isolators



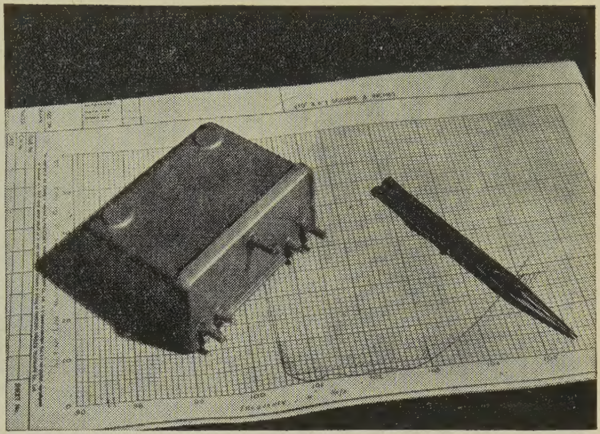
Existing Marconi designs cover most of the frequencies from 1.8 Kmc/s to 10 Kmc/s, and enable the following performance characteristics to be obtained:
Reverse losses of 45 dB or greater VSWRs 1.03.
Forward losses of 0.5 dB or less.
A wide range of ferrite devices, including circulators, is available; if your requirements fall outside our existing range we shall be pleased to consider a design to suit you.

50 ohm Coaxial Connectors



Designed to combine low VSWR with positive mechanical connection at 5000 mc/s, this connector is unique in its use for laboratory measurements and for any application calling for a stable and reliable match. Alternative arrangements are available for flexible cables or panel mounting use. The connectors are normally female but can be converted to male by the simple addition of a locating ring and connecting pin. The design incorporates a sealing ring for waterproofing the connecting joint.

Miniature Crystal Filters



Crystal filters operating at 100 kc/s with bandpass characteristics.

Pairs of upper and lower sideband selection filters for SSB applications.

IF filters at 455 kc/s.

Marconi's are constantly engaged in the development of new crystal filter techniques, and it is impossible to specify the large number of filter designs available. The Company's unique experience in this field enables us to offer advice and technical assistance with filter problems of any kind.

MARCONI

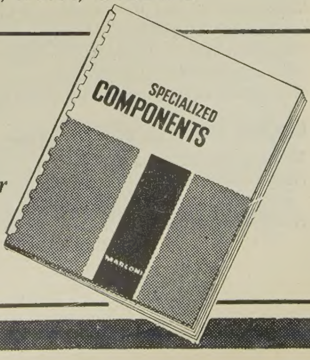
SPECIALIZED RADIO COMPONENTS

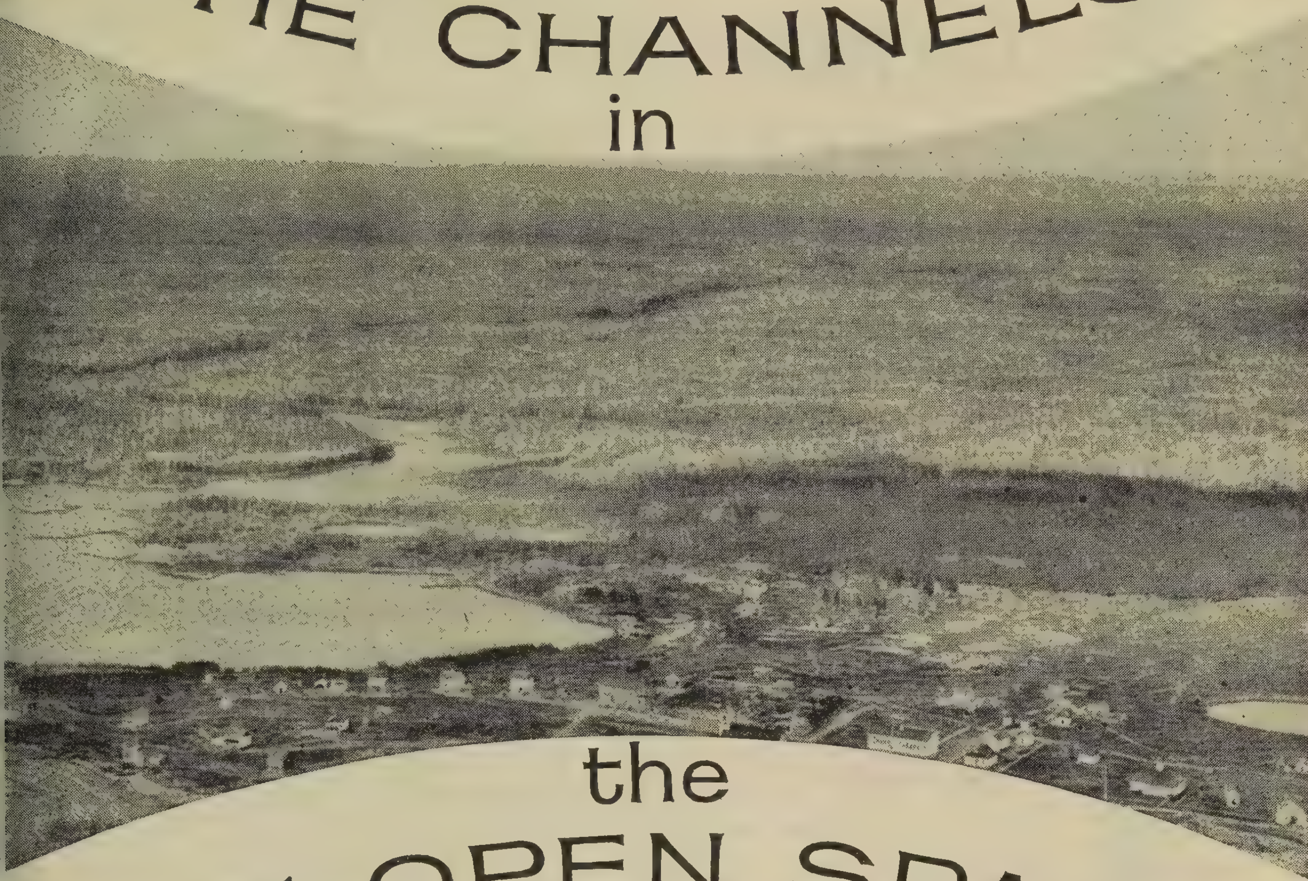
Write for details of crystal filters and other specialized components to:

SPECIALIZED COMPONENTS GROUP

MARCONI'S WIRELESS TELEGRAPH COMPANY LIMITED, CHELMSFORD, ESSEX, ENGLAND

A preliminary catalogue is available listing specialized components for immediate delivery.





EI focus on . . .

RMANIUM GOLD-BONDED DIODES

MPUTERS, DATA PROCESSING and INSTRUMENTATION APPLICATIONS and for
AL LOW-POWER RECTIFICATION where high efficiency is required

ADVANCE INFORMATION

TANDING FEATURES * LOW FORWARD RESISTANCE HENCE HIGH EFFICIENCY
NGE OF P.I.V. AVAILABLE * SUB-MINIATURE ALL-GLASS CONSTRUCTION * HERMETIC SEAL

LD-BONDED DIODE—Type CG84H

Inverse Voltage:	12 volts
ard Voltage drop at 40 mA, 25°C:	0.6 volts
se Current at –12V, 25°C:	12 microamperes
at –12V, 60°C:	38 microamperes
Forward Current at 25°C:	90 milliamperes
ion Capacitance at –1V:	2.4 picofarads
d Charge at –10V inverse, ard and 100 ohms circuit impedance:	0.6 milli-micro-coulombs

D GOLD-BONDED DIODES—
80H, CG81H, CG82H, CG83H and CG85H

Ratings

bient temperature	CG80H	CG81H	CG82H	CG83H	CG85H
se voltage V	100	75	50	25	12
ard current mA	70	75	77	80	90
aged over any 20 ms period) mW	100	110	110	110	114

Characteristics

bient temperature		CG80H to CG83H		
urrent	mA	10	30	100
op. Typical	V	0.375	0.50	0.8
Maximum	V	0.45	0.65	1.1

		CG85H
urrent	mA	40
op. Maximum	V	0.6

haracteristics		CG80H	CG81H	CG82H	CG83H	CG85H
	V	100	75	50	25	12
erse						
25°C	μA	30	38	20	12	—
60°C	μA	150	180	120	75	—
nverse						
25°C	μA	100	75	50	25	12
60°C	μA	500	300	250	150	38

ce Typical production spreads measured at –1 volt are as follows: Minimum 1.1 pF; Typical 1.8 pF; Maximum 3.0 pF.

ge When measured with 10 mA forward current, 10 volts inverse 100 ohms circuit impedance hole storage values are as follows: mμC; Maximum 3.0 mμC.

FEATURES—

common to all types (CG80H to CG85H)

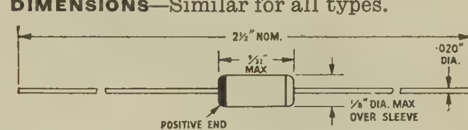
ENCAPSULATION All these AEI Gold-bonded Diodes are of all-glass construction with lead-in wires sealed into the encapsulating tube. The expansion of the wire is matched to that of the glass, providing a hermetic glass-to-metal seal over a wide range of ambient temperatures. The junction is thus fully protected and the characteristics of the diode are very stable because no fluxes, varnishes or other harmful internal contaminants are used.

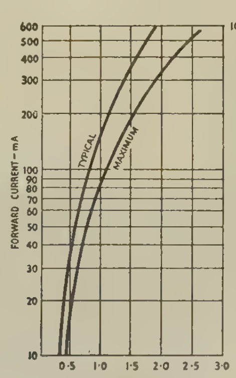
SHOCK-RESISTANCE Because of their small size (about ⅛ in. diameter and ¼ in. long), and the rigidity of the small component assemblies, these diodes are able to withstand severe mechanical vibration and shock.

SILVER-PLATED LEADS The lead-in wires are silver-plated to facilitate soldering, but a thermal shunt should be used if heat is applied close to the diode body. In any case, joints should not be made closer than ¼ in. to the diode body and the leads should not be bent closer than 1/16 in.

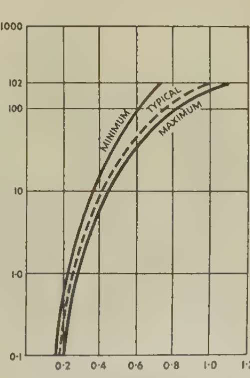
IDENTIFICATION Normally each type of diode is identified by a plastic sleeve and the polarity is shown by marking the cathode end of the body red.

DIMENSIONS—Similar for all types.





Forward characteristic CG80H to CG83H

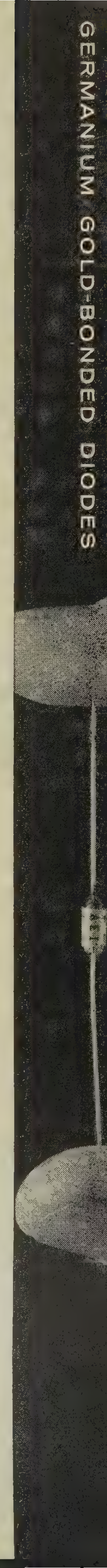


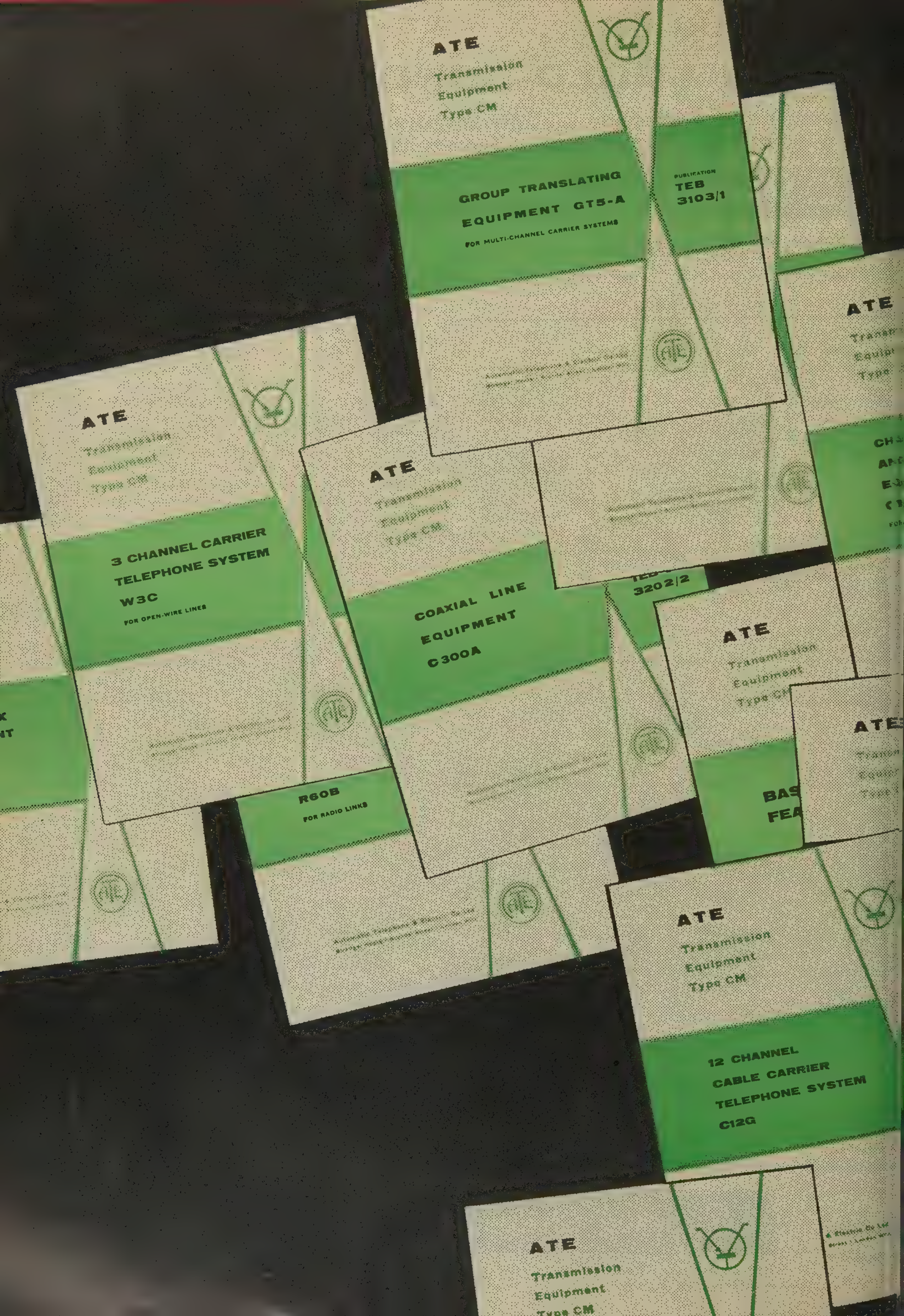
Forward characteristic CG84H to CG85H

AEI Associated Electrical Industries Limited

ELECTRONIC APPARATUS DIVISION VALVE & SEMI-CONDUCTOR SALES DEPT
CARHOLME ROAD, LINCOLN. TEL: LINCOLN 26435

A5654



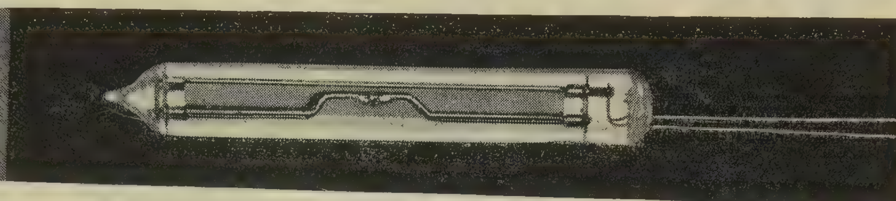

ATE
Transmission
Equipment
Type CM
GROUP TRANSLATING
EQUIPMENT GT5-A
FOR MULTI-CHANNEL CARRIER SYSTEMS
PUBLICATION
TEB
3103/1
ATE
Transmission
Equipment
Type CM
3 CHANNEL CARRIER
TELEPHONE SYSTEM
W3C
FOR OPEN-WIRE LINES
ATE
Transmission
Equipment
Type CM
COAXIAL LINE
EQUIPMENT
C300A
3202/2
R60B
FOR RADIO LINKS
ATE
Transmission
Equipment
ATE
Transmission
Equipment
Type CM
12 CHANNEL
CABLE CARRIER
TELEPHONE SYSTEM
C12G
ATE
Transmission
Equipment
Type CM

NEW RANGE OF

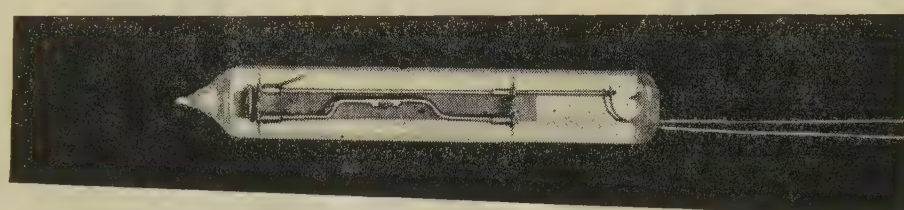
miniature crystals

TO REPLACE STC Type 4407 (55–83 kc/s) and STC Type 4028 (75–140 kc/s)

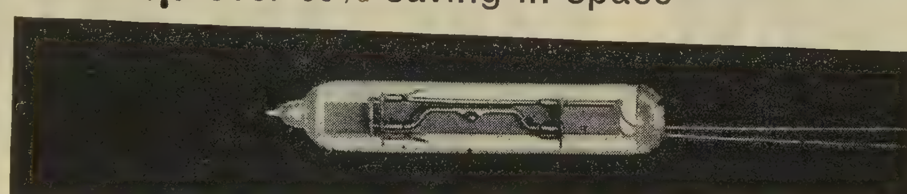
STC



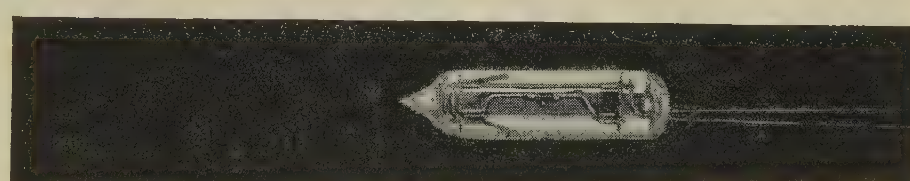
* only 0·4 in. overall diameter



* over 60% saving in space



* ideal for horizontal mounting on transistorised printed circuits



Illustrations are full size

Frequency Range kc/s	STC Type	Maximum seated height		Diameter		Volume		Lead length	
		in.	cm	in.	cm	cu. in.	cu. cm	in.	cm
57 – 62·9	4432	2·83	7,19	0·4	1,02	0·36	5,90	1·5	3,8
63 – 71·9	4438	2·44	6,19	0·4	1,02	0·31	5,08	1·5	3,8
72 – 99·9	4437	2·04	5,18	0·4	1,02	0·26	4,26	1·5	3,8
100 – 150	4435	1·45	3,68	0·4	1,02	0·18	2,95	1·5	3,8

Write for further information to :—

STC COMPONENTS GROUP

Standard Telephones and Cables Limited

QUARTZ CRYSTAL DIVISION · HARLOW · ESSEX

61/4MQ

• FOR −40°C TO +100°C OPERATION
• HIGH CAPACITANCE STABILITY
• CONFORMING TO CATEGORY H1 of RCS131 and BS2131
STC
HIGH GRADE
PAPER CAPACITORS
OIL IMPREGNATED · OIL FILLED
• HIGH INSULANCE
• LONG LIFE
TUBULAR METAL CASES
Capacitance Range: 0.0033μF to 1.50μF
Voltage Range: 100V to 350V d.c. 75V to 250V a.c.
Seals: Sintered glass compression type
RECTANGULAR METAL CASES
Capacitance Range: 0.25μF to 10.0μF
Voltage Range: 150V to 1 500V d.c. 75V to 750V a.c.
Seals: Corundite
Other STC paper capacitors include miniature paper tubular and a wide range of paper capacitors conforming to British Post Office specifications. Full details are available on request.

PAPERS FOR THE PROCEEDINGS

Handbook for Authors — Anyone who is thinking of submitting a Paper to The Institution should apply to the Secretary for a copy of the **Handbook for Authors**. The price is 3s. (post free), but a copy will be supplied free of charge if the application is accompanied by a summary of the Paper. The following are some of the main points considered in the Handbook.

Acceptability — To be acceptable, a Paper should normally contribute to the advancement of electrical science or technology. The Institution does not accept Papers which have been published elsewhere.

Length — No Paper should occupy more than 10 pages in the *Proceedings*. Authors can generally keep well within this limit. For example, the average Paper published in 1960 consisted of 6 000 words (5 pages) and, with its illustrations and mathematics, occupied a total of 8 pages.

Summary — An essential part of a Paper is the **Summary**, which should not exceed 200 words.

Text — The **Text** should begin with sufficient introductory matter to enable the Paper to be understood without undue reference to other publications.

The Text should include no more mathematics than is essential. Extended mathematical treatment and lengthy digressions—if they must be included—should be put in **Appendices.**

Proprietary articles should not be mentioned by name unless this is unavoidable.

The rationalized M.K.S. system of units is preferred.

Acknowledgments — Assistance in the preparation of the Paper, and sources of information, should be acknowledged. References to manufacturers should be made only under **Acknowledgments.**

References — Bibliographical **References** should be numbered and listed in a special section, and indicated in the Text by means of 'indices'.

Numbering — The Text should be appropriately sectionalized, the sections and their subdivisions being numbered according to the 'decimal' system. Acknowledgments, References and Appendices should be numbered as though they were sections of the Text.

Typing — Typing should be on one side of the paper only, with double spacing between lines and a $1\frac{1}{2}$ inch margin on the left. Besides the original typescript, two carbon copies are required by The Institution.

Advice on the typing of mathematics is given in the **Handbook for Authors,** which includes a facsimile of a typewritten page containing mathematics.

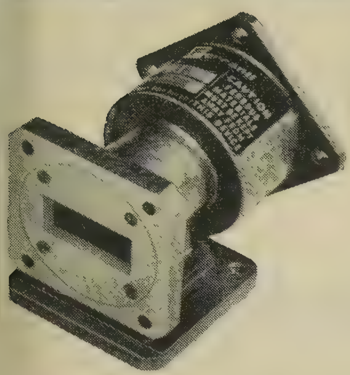
Illustrations should not be drawn or pasted on the typewritten pages. They are of no use to the printer, but he does need a complete list of captions, again with double spacing. The list should be attached to the typescript.

Drawings — Three sets of drawings, which may be in the form of dye-line prints, should accompany the typescript. Tracings, which will be required later, should be in indian ink with the lettering in pencil. The reduction in the size of the drawings, and therefore the size of lettering required, is not settled before the Paper has been accepted.

Dispatch — The typescript and illustrations should be packed flat, not rolled, and addressed to *The Secretary, The Institution of Electrical Engineers, Savoy Place, London, W.C.2.*

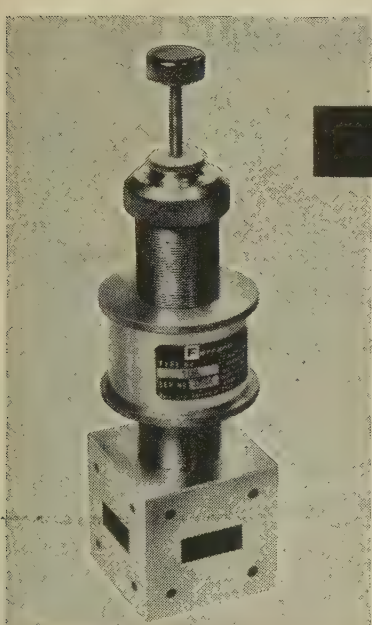
FERRANTI

Ferrite Circulators



X-BAND 3-PORT FERRITE CIRULATORS TYPES 5M/16.1 AND 5M/16.2A

	5M/16·1	5M/16·2A
Bandwidth	9375 ± 100 Mc/s	8800 ± 30 Mc/s
Temperature Range	—40 to + 80°C	—40 to +80°C
Power Rating	50kW peak 50 W mean	50 kW peak 50 W mean
Weight	12 oz	6·5 oz



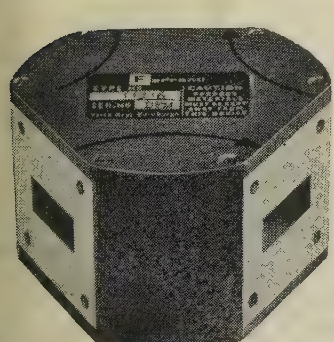
TUNABLE 4-PORT FERRITE CIRCULATORS SERIES 10A

Tuning Range..................8000–9600 Mc/s

Power ....20C W

Weight...........................4½ lb

Performance at Centre Frequency		Performance for a bandwith of ± 30 Mc/s	
Isolation	> 30 db	Isolation	> 15 db
Insertion loss	<·3 db	Loss	< 0·4 db
V.S.W.R.	< 1·25	V.S.W.R.	< 1·25



3-PORT FERRITE CIRCULATORS SERIES 11M

Frequency ....Centred in the band 8000-11000 Mc/s

Bandwidth	Loss	Isolation	V.S.W.R.
	0·10 db	> 35 db	1·05
100 Mc/s	0·15 db	30 db	1·08
200 Mc/s	0·20 db	25 db	1·12
400 Mc/s	0·25 db	20 db	1·20

Max. Power ....50 kW. Peak
1kW. Mean

Weight..............1½ lb approx.

FERRANTI

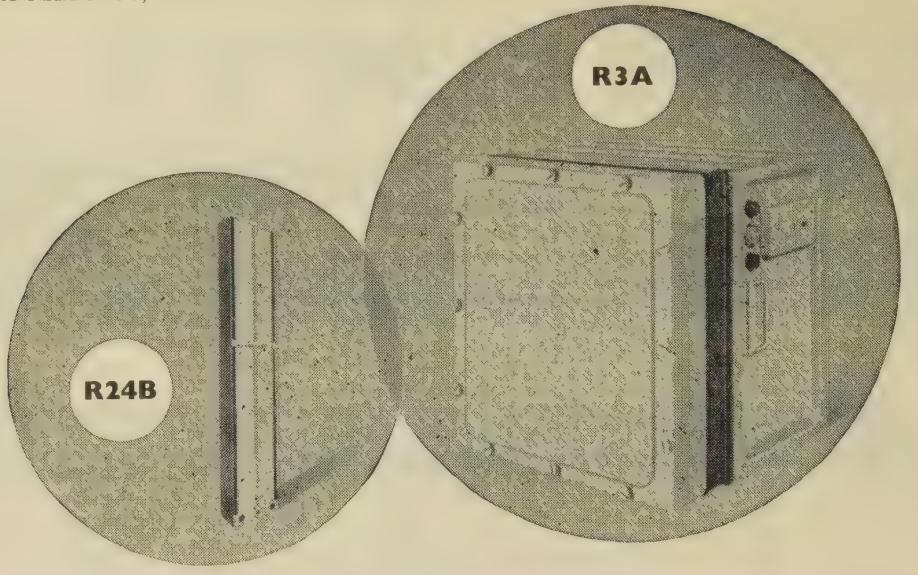
First into the Future

Write for further details to:

FERRANTI LTD · KING'S CROSS ROAD · DUNDEE

Tel: DUNDEE 87141

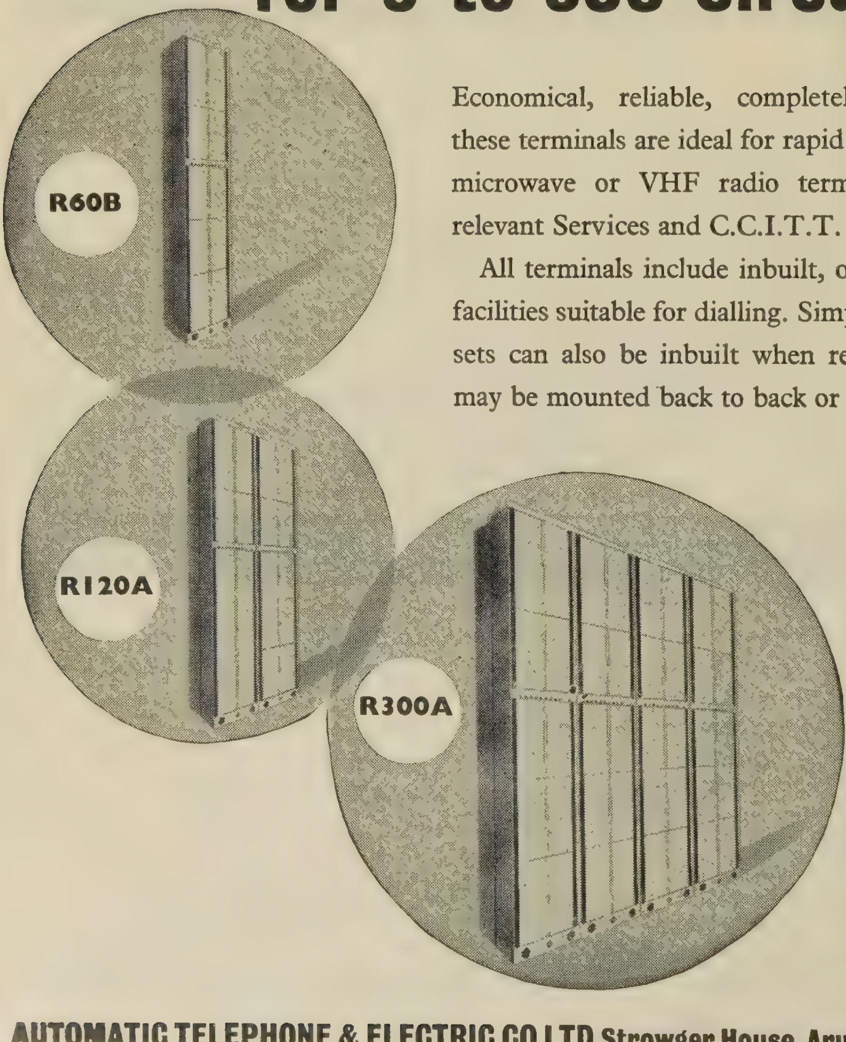
DS/T 72



R3A A trans- portable 3 chann terminal designe for field or Servi use. Write for Bulletin TEB. 33

R24B A 12/24 channel terminal in the 6-108 Kc/ spectrum, compl on one rackside. Write for Bulle TEB. 3301

A.T.E "Packaged" Radio Channelling in Fully-transistorised Basic Units for 3 to 300 Circuits



Economical, reliable, completely self-contained, these terminals are ideal for rapid installation at any microwave or VHF radio terminal. They meet relevant Services and C.C.I.T.T. requirements.

All terminals include inbuilt, outband, signalling facilities suitable for dialling. Simple ringdown relay sets can also be inbuilt when required. Racksides may be mounted back to back or side by side.

R60B A 60 channel terminal (12-252 or 60-300 Kc/s), complete on two racksides. Write Bulletin TEB. 33

R120A A 120 channel terminal (12-552 Kc/s), complete on four racksides. Write Bulletin TEB. 3

R300A A 30 channel terminal (60-1300 Kc/s), complete on eigh racksides. Write Bulletin TEB. 3

TRANSMISSIO EQUIPMENT TYPE CM for li cable and radio systems

AUTOMATIC TELEPHONE & ELECTRIC CO LTD Strowger House, Arundel St., London, WC2. Tel. Temple Bar 9

INDEX OF ADVERTISERS

THE PROCEEDINGS OF
HE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

L. 108. PART B. No. 40. JULY 1961

391.812.624: 621.396.677

The Institution of Electrical Engineers
Paper No. 3576 E
July 1961

TROPOSPHERIC SCATTER OBSERVATIONS AT 3480 Mc/s WITH AERIALS OF VARIABLE SPACING

By R. W. MEADOWS, B.Sc.(Eng.), Associate Member.

(*The paper was first received* 9*th November*, 1960, *and in revised form* 17*th February*, 1961.)

SUMMARY

ıort-term measurements are described of signal amplitude, fading acteristics, correlation distances and height gain at three ranges , 258 and 398 km) from a transmitting system radiating horially polarized waves. Although the results were variable and at s difficult to interpret, it is considered that 150λ should be an ıuate horizontal spacing, even at the shortest ranges, for directional iving aerials operated in diversity, with a tendency towards smaller es for vertical spacing. Fading rates lying generally between 1 10 per second are found, with Rayleigh-type fading tending to ır at the longer ranges. No consistent relationship between signal litude, fading and correlation distance is found, although the est fading rates are often accompanied by low correlation distances. ss-correlation measurements show definite evidence of the steady ing of scattering centres. Diffraction effects leading to height are found, and there is some indication of standing waves due to ınd reflections. Aerial siting is discussed and the transportable pment which was developed for the tests is briefly described.

(1) INTRODUCTION

he weak, rapidly fading signal arriving at a receiving point ı a transmitter beyond the radio horizon is often conred[1, 2] to result from energy scattered from all parts of the osphere common to both the transmitter and receiver beams. energy incident at the receiver is therefore spread out over ıngle.[3] It is well known that the instantaneous amplitude of ı a signal at a point in space varies with the position of that ıt, so that the outputs from a pair of receiving aerials, spaced t transversely to the path, do not fade identically. The way hich they differ over a period can be expressed in terms of ımplitude correlation coefficient, r, and the way in which r eases from unity to zero as the spacing, s, is increased from to infinity can be used to estimate the angular spread of the lent energy, provided that certain assumptions about the structure of the latter are made.[3, 4] In particular, if both hape of the r/s curve and that of the angular energy distribution measured in the same plane are Gaussian, it can be shown that most of the energy in that plane is contained within an angle of approximately $30\sqrt{(1-r)}/(s/\lambda)$ degrees. Also, the value of s at which r has reduced substantially to zero can be taken as a reasonable value for the spacing between two receiving aerials when operated in a diversity system. A more precise way of indicating the width of the r–s function (especially when it is Gaussian) is in terms of the correlation distance,[5] which is the value of s at which $r = 1/\varepsilon$.

It is clear that a knowledge of the angular spread of the energy, together with the correlation distances, amplitude and fading characteristics, is of great importance in the design of tropospheric-scatter diversity systems, and may also throw some light on the nature of the scattering mechanism. As part of a programme of work designed to study these matters, a transmitter was installed at Start Point in Devon, and it was decided to construct a transportable spaced-aerial twin-channel receiving system which could be used by the Radio Research Station to supplement the work already being carried out at fixed receiving stations by other organizations.[6, 7] The system is described briefly in this paper. The aerials could be separated by any value from zero to about 70λ horizontally, or 200λ vertically. An analogue computer was constructed to enable r to be estimated readily. At the same time it was felt desirable to make pen-recordings of the mean level of the signal amplitude, and (when possible) to record samples of the instantaneous signal amplitudes on film and tape for later processing to derive the auto- and cross-correlation functions. The aerial system was also suitable for measuring variations of mean field strength occurring with height and plan position.

Measurements at three distances (130, 258 and 398 km) from the transmitter were made. A distance of 86·5 km in the transmission path at the transmitter end was over sea. The time spent at each receiving point was 2–3 days per week over a period of at least three months, which was sufficient to enable a fair idea of the short-period variations to be obtained. During most of this time a receiver was also in operation in a fixed position at Wembley,[6] which served as a useful control. The paper presents a general account of the measurements made and a discussion of

itten contributions on papers published without being read at meetings are d for consideration with a view to publication.
e paper is an official communication from the Radio Research Station, Depart- of Scientific and Industrial Research.

their implications, rather than a detailed analysis of all the results obtained.

(2) APPARATUS

(2.1) Transmitting System

The transmitting system was installed[6] near the edge of a 126m cliff at Start Point, Devon, and beamed towards the receivers on a true bearing of 056°. A specially designed cavity magnetron was used, delivering a continuous-wave output of 500W to an 8ft (28λ) diameter paraboloid reflector producing horizontally polarized radiation within a beam of width 2·6° at 3dB points. The plane-wave aerial gain relative to an omnidirectional radiator was 36·5dB, so that the energy was concentrated into a 'beam-power' of 2·2MW.

(2.2) Twin-Channel Spaced-Aerial Transportable Receiving System

(2.2.1) Receiver and Amplitude-Recording Arrangements.

Two identical receivers were used, one connected to each aerial (see Section 2.2.4). The receivers were completely separate and similar to those described in Reference 6 except that the first and second beating oscillators were common to both receivers. The first i.f. was 31Mc/s and the second was 1·6Mc/s. The noise factor of each was 11dB and the final bandwidth 60kc/s, except when narrowed to 12kc/s on one occasion to increase the signal-noise ratio. Automatic frequency control was applied in two stages: coarse control was effected by a voltage applied to the reflector of the klystron used as the first beating oscillator, and fine control, by means of a reactance valve coupled to the second oscillator. The output of each receiver at 1·6Mc/s was rectified, smoothed and applied to a pen recorder, the overall time-constant being about 6sec. The two outputs were further amplified at 1·6Mc/s and, after high-level rectification, applied through 200c/s low-pass filters to the Y1 and Y2 plates respectively of a double-beam cathode-ray tube, so that the rapid signal fluctuations in the two channels corresponding to the spaced aerials could be readily compared and photographed side by side on a single strip of 35mm film.

(2.2.2) Recording of the Correlation between the Rapid Fluctuations.

The two high-level d.c. fluctuating outputs (or channels) corresponding to the spaced aerials were also fed to a specially constructed correlator.[8] This equipment provided a pair of continuous pen-recordings enabling a 'running value' of r to be very quickly estimated as follows. One pen recorded the value of the square of the instantaneous sum of the fluctuations in each channel about a mean level of zero at the same time as the other recorded the square of the difference; r was then given by the difference between corresponding readings, divided by the sum. The behaviour of the correlator was occasionally checked from the film records of instantaneous amplitude, and for this purpose it was found that a period of about 1min with at least a few hundred pairs of samples was adequate to enable r to be calculated to an accuracy of $\pm 0{\cdot}1$.

In the correlator it was necessary to translate the fluctuations occurring in each signal channel to a mean value of zero in order that the principle of operation of the device should be valid. This was done by passing each rectified signal-voltage through a *CR* combination. The time-constant of this had to be long enough to ensure that the peaks and troughs of the fading signal were not substantially clipped, yet short enough to reject drifts of mean level due to long-period propagational changes. A time-constant of 90sec was ultimately chosen, but this could usually be reduced to 9sec without appreciable change in the character of the recordings. A disadvantage of a long time-constant was that the equipment had to be run for a consider time before useful recordings could be made, to allow the m level of the fluctuations to decay sufficiently close to zero addition, it was soon found that the large and frequently oc ring fluctuations due to reflections from aircraft produced sistent effects if the time-constant were too great.

It was found by experience that the pen-recordings coul quickly smoothed and averaged by eye over a period of long enough to give the correlation coefficient sufficie accurately, spurious peaks obviously due to aircraft b ignored in the process. It was not considered worth whil provide further equipment enabling the correlation coeffic to be recorded directly; in fact, this was not altogether desira since such unqualified results might have been misleading ov to the presence of aircraft or to other spurious effects.

(2.2.3) Production of Signal Auto-Correlation and Cross-Correlati Functions.

The auto-correlation function of the fast-fading signa either channel is important since it contains much informa about the signal variability, and after suitable manipulation g the frequency spectrum and fading rate. The correlation co cient directly between the outputs of the two channels has alre been dealt with, but if a time delay is inserted in one chan new value of r is obtained leading to the cross-correlation f tion as a plot of r against delay. The use of these two kin function is considered in Section 5.4.

The derivation of either function requires that the sig shall first be recorded. The fluctuating d.c. signal in channel was therefore arranged to modulate the amplitude 1·5kc/s carrier in an accurately linear manner giving carrier level at zero signal. The two modulated carriers then recorded separately and simultaneously in a twin-cha magnetic-tape recorder. The output of the recorder coul amplified, demodulated, and applied when necessary to a recorder, the Y-plates of a cathode-ray tube, or to the corre input. The tape recorder had been modified by looping tape between the two spaced recording heads so that a delay having any value between 0 and ± 2sec could be introd into one channel with respect to the other. Consequently tape recorder and correlator formed a unit enabling correlation and cross-correlation functions to be obtained. equipment was not completed until the last stage of the pr experimental work, and was therefore used only for the obtained at the final site.

(2.2.4) Aerial and Mast System.

This T-shaped structure is shown in Fig. 1. The ve member consisted of five triangular lattice sections 12½ft giving a total height of 62½ft. The horizontal member con of two similar sections, and the whole structure was supp rigidly by four stays attached to the ends of these sec Carriages supporting aerials were arranged to run along the vertical or the horizontal members, and could be m either by hand-winches on the ground or by a slow-m electric drive. When the latter was used, the carriages cou made to reverse their directions automatically when extren travel were reached.

The aerials used were either (*a*) 1·25m (13λ) diameter boloids with 120° horn feeds, as shown, or (*b*) small 2λ-s pyramidal horns. Both were arranged for horizontal pol tion. The measured plane-wave gain of each paraboloi 29dB (relative to an isotropic radiator), and the beam 5° between 3dB points; the corresponding figures for the were 17dB and 32°. Mounted with each aerial was a rin mixer and preamplifier delivering i.f. signals at 31Mc/s

Fig. 1.—Transportable spaced-aerial system.

ble coaxial cable running down the mast. The associated tron oscillator was in a trailer housing the rest of the equip-t, the output from it being taken up the mast by a single ble coaxial cable and thence to the mixers via a T-connector. r flexible coaxial cables enabled standard i.f. signals to be ied to each preamplifier for calibration purposes. The inter-n between the two systems was less than −30 dB, even at mum aerial spacing.

(3) MEASUREMENTS AT SLOUGH (258 km)

ie first measurements were made at the Radio Research on (15 m above sea level) primarily to test the whole system re transporting it elsewhere. The site is not ideal since local subtend an elevation angle of $\frac{1}{2}°$ in the direction of the smitter. Had the horizon been unobstructed the angular nce[9] would have been 32 mrad, assuming an effective radius e earth of 8 000 km, but the actual value was 40. Measure-ts were made at intervals between August, 1957, and ch, 1958.

) Correlation between the Fast Fading in the Outputs of Two Spaced Receivers

) Vertical Aerial Spacing.

ie paraboloid receiving aerials were used first, one being fixed position at the top of the mast and the other moved nd down. It was soon found that, even with the aerials cent, the correlation coefficient did not often rise to unity, ie paraboloids were replaced by the horns, thus enabling ings as small as 2λ to be obtained. The spacing was ed in steps of about 4λ, a minute or two being spent at each position to allow a significant value of the correlation coefficient to be obtained. The results were found to be very variable, the correlation coefficient falling substantially to zero at spacings varying from 10λ up to a maximum of about 30λ. Some specimen curves are given in Fig. 2, and it will be seen that they

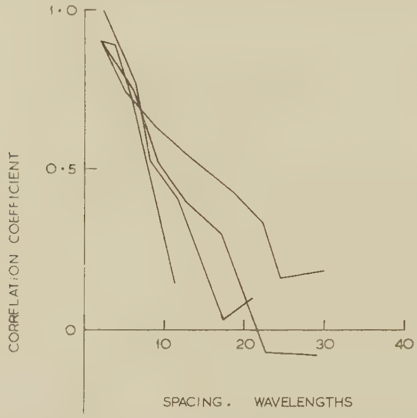


Fig. 2.—Variation of correlation coefficient with vertical aerial spacing (Slough).

are not smooth; this may have been due to changes in propagation conditions while the experiments were in progress, which is supported by the fact that when the tests were repeated as quickly as possible the irregularities were not usually repeated. Nevertheless, it is interesting to attempt to compare the results obtained with those to be expected from a simple theory of turbulent scattering leading to a Gaussian-like *r–s* function, although it is now realized that such a theory does not necessarily describe conditions completely. Gordon has shown[5] that the spacing, D_v, required for the correlation coefficient to fall to $1/\varepsilon$ (correlation distance) is given by $\lambda a/d$, where a is the effective radius of the earth and d, the path length, an assumption being that the transmitter beam width is broad enough to ensure that the scatterers which produce substantial amounts of energy at the receiver are all illuminated. The energy then arrives at each aerial over a vertical angle of $d/2a$ radians between half-power points, which assumes a doubling of the direct-wave spread owing to reflection at the ground. In this instance the angle becomes 1·2°, which is within the width of the beam of the transmitting aerial if the scattering is assumed to occur mainly near the mid-path; consequently it is likely that the transmitter beam width is broad enough for the previously mentioned assumption to be valid. In this case, D_v becomes 24λ. The experimental curves are very irregular, but if approximating Gaussian curves are drawn through them, the correlation distances estimated from these will be found to encompass the theoretical figure of 24λ. The whole range of these correlation distances estimated in this way implies that the energy spreads over about 1–3° in the vertical plane, depending upon atmospheric conditions. It would have been better to derive the angular distribution of energy corresponding to each experimental *r–s* function by a formal process of integration.[3] This was not done because the irregularities in the curves, obtained from successive runs taken as quickly as possible, did not repeat, and it was therefore considered that the simpler process adopted was no less justifiable.

(3.1.2) Horizontal Aerial Spacing.

The tests described in Section 3.1.1 were repeated with horizontal instead of vertical spacings, and the results were again variable, the correlation coefficient falling substantially to zero at

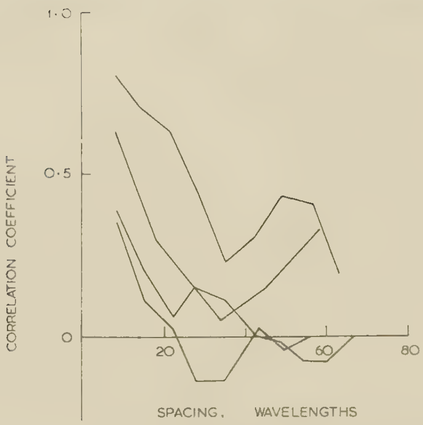


Fig. 3.—Variation of correlation coefficient with horizontal aerial spacing (Slough).

spacings of 20–60λ, some specimen curves being shown in Fig. 3. Gordon's scattering theory now gives the spacing, D_h, for the correlation to fall to $1/\varepsilon$, as $3\lambda/4d$, which is smaller than D_v and becomes 18λ for this path. This corresponds to energy arriving (between half-power points) over a horizontal angle of $2d/3a$, compared with $d/2a$ for the vertical case, and becomes 1·6°. However, approximating Gaussian curves drawn through the experimental results shown in Fig. 3 indicate that D_h is greater than D_v, whereas theory shows the reverse. This is discussed in Section 6. If it is assumed that the theoretical correlation distance of 18λ corresponds to an angular spread of 1·6° in the practical case, then 20–60λ corresponds to an angular spread of 1·6–0·5° for the incident energy in the horizontal plane, which is about half that estimated for the vertical.

In addition to these variable-spacing measurements, a number of measurements were made with the pair of paraboloids operated at a fixed horizontal spacing of 70λ. The results obtained are shown in Fig. 4. It can be seen that r varied between about 0 and 1, although the latter figure was rare. A feature of the results was that r tended to stay around a particular value for hours rather than minutes. The period 19th–21st November, 1957, is interesting because, during a country-wide spell of persistent fog, exceptionally slow fading was experienced (on the 20th) with high correlation. Furthermore, low field strength was obtained, and a diversity system would therefore have given little improvement unless the spacing between the receiving aerials was much greater than 70λ. The aerials were at a height of 19 m, and it is just conceivable that during these exceptional weather conditions they were above a low duct of the type observed at sea,[10] leading to the low field strength observed. Height-gain measurements would have been interesting in this context, but unfortunately were not possible.

(3.2) Height-Gain Measurements

For height-gain tests the paraboloids were used. One was maintained in a fixed position at the top of the mast as a control. The other was wound very slowly up and down, a complete cycle occupying about 15 min, so that the amplitude variations due to the fast fading could be smoothed out. As a result of these tests it could be said that a consistent 5 dB height gain was obtained, occurring mostly in the first 6 m. This suggests diffraction around some fairly local obstacle; according to standard knife-edge diffraction theory a row of houses 1·5 km away could have accounted for the effect. Ripples were occasionally present on the height-gain curve, and this effect is mentioned in more detail in Sections 4.2 and 5.2, where the presence of much more consistently occurring ripples at other sites is attributed to a downward ground slope in front of the receiving aerial system. However, no such ground slope exists at Slough, and the ripples which occasionally occurred must be attributed to some other cause such as interference between the direct and reflected components of a single stable downcoming ray.

(3.3) Amplitude and Fading Measurements

Pen-recordings (with a time-constant of 6 sec) of mean amplitude were made whenever the equipment was running, and hourly samples of the fast fading were taken on film. To enable absolute field strength to be deduced from the records, calibrating signals were provided by a 3480 Mc/s oscillator about 0·4 km away.

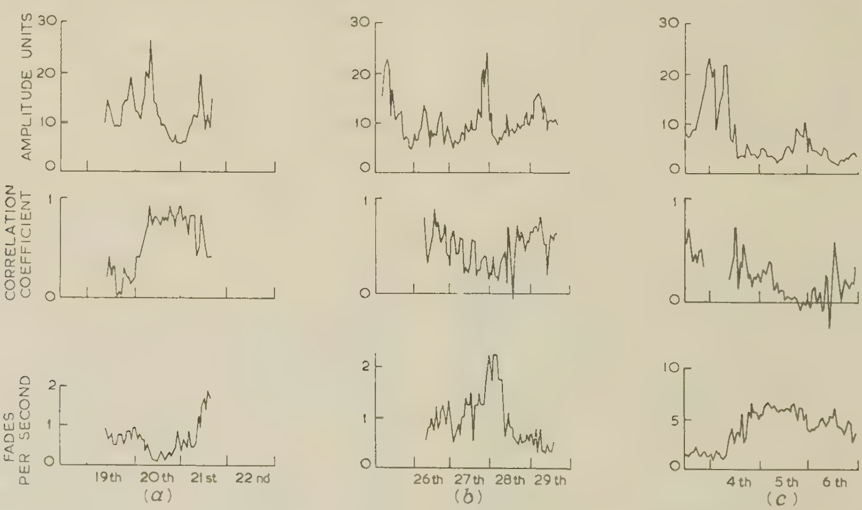


Fig. 4.—Variation of amplitude, correlation coefficient and fading rate with time (Slough).
(*a*) 19th–21st November, 1957. (*b*) 25th–29th November, 1957. (*c*) 3rd–6th February, 1958.

The measurements were rather intermittent, but a mean value of the transmission loss is given in Table 1 for each month when measurements were made during 1957–1958. The values include measurements taken at all times of day, since no significant diurnal variation was found. The transmission loss is defined as

Table 1

MONTHLY MEAN TRANSMISSION LOSS

Month (1957–1958)	Measured transmission loss (dB)	Derived basic path attenuation (dB)
August ..	149	213·5
September ..	151	215·5
October ..	146	210·5
November ..	149	213·5
December ..	152	216·5
January ..	154	218·5
February ..	150	214·5

Nominal gain of transmitting aerial = 36·5 dB.
Nominal gain of receiving aerial = 29 dB.

the power input to the transmitting aerial divided by the output from the receiving aerial. The variability was similar to that obtained for Wembley.[7]

The figures in the right-hand column are calculated from the monthly measurements assuming (*a*) 0 dB gain for both the transmitting and receiving aerials, and (*b*) that a total aerial-medium coupling loss of 1 dB was present. This figure was estimated from the mean horizontal and vertical spreading of the incident energy which was found in the results described in Section 3.1. The figures for the nominal gains of the aerials were the measured plane-wave gains referred to an isotropic radiator.

The fading rate over a period was defined[4] as the number of upward crossings per second through the mean signal level averaged over that period. The period of the film samples was about 1 min. The fading rate normally lay between 1 and 10 but occasionally fell below 1. An attempt was made to correlate the fading rate with the transverse component of wind speed in the common volume, but the nearest meteorological stations were too far away and the wind-speed measurements too infrequent for sufficiently accurate data to be obtained. However, a definite trend was observed; the higher the wind speed, the faster the fading, as might be expected.[4] This confirms similar measurements made at Wembley[6] over this path.

There was also a tendency for fast fading (greater than about 5 upward crossings per second) to be accompanied by low correlation distance in both the vertical and the horizontal directions, but, on the other hand, slow fading was not necessarily accompanied by high correlation distance. This has been embodied in a chart shown in Fig. 5.

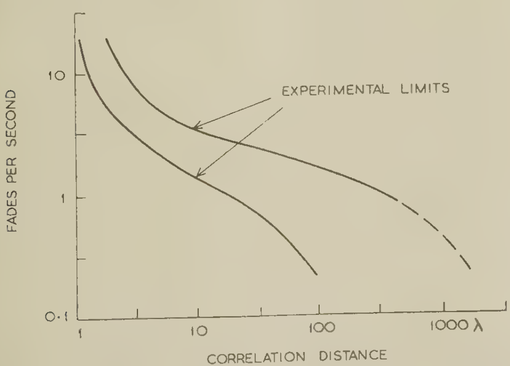


Fig. 5.—Variation of fading rate with correlation distance.

(4) MEASUREMENTS AT IPSWICH (398 km)

The site was near Witnesham, about 5 miles north of Ipswich, and at a height of 53 m above sea-level. In the foreground the land fell gently away and rose again to just below the height of the site in a total distance of about 4 km. The horizon was just obscured by the tops of a thick belt of trees, the angular distance being 50 mrad. The measurements were made for about three days per week during April–July, 1958. Another receiver was being operated nearby on long-term field-strength recording trials.[7]

(4.1) Correlation between the Fast Fading in the Outputs of Two Spaced Receivers

(4.1.1) Vertical Aerial Spacing.

The results and their variability were similar to those at Slough (Section 3.1.1) but with a tendency for the correlation distances to be less. The signals at this range were too weak for the small horns to be used (even though the receiver bandwidth was reduced from 60 kc/s to 12 kc/s), so all measurements were made with the paraboloids (13λ in diameter).

(4.1.2) Horizontal Aerial Spacing.

Several runs were made to determine the variation of correlation coefficient with spacing, the paraboloids being separated or drawn together continuously at a speed of 1·57 ft/min. However, whenever such a horizontal spacing run was attempted the correlation coefficient, even at the minimum spacing obtainable (17λ), never rose significantly above zero. The aerials were therefore left at this spacing and *r* was recorded continuously. Most of the time it did not exceed 0·5, but during one exceptional spell it did in fact approach unity during two successive nights when the fading rate was exceptionally low, as illustrated in Fig. 6. For the next few nights, therefore, some spacing runs

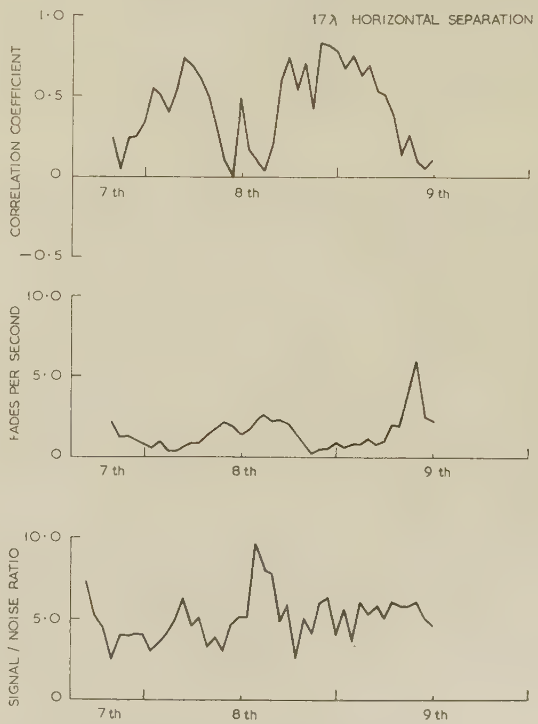


Fig. 6.—Variation of correlation coefficient, fading rate and signal-to-noise ratio with time (Ipswich).
7th–9th May, 1958.

were attempted, but the phenomenon did not repeat and no results showing a significant variation of *r* with spacing were ever obtained. Consequently it is inferred that the spacings

required for the outputs to be substantially uncorrelated were usually considerably less than the 20–60λ found at Slough, and therefore that the angular spread was somewhat greater than the 0·5–1·5° range found there.

(4.2) Height-Gain and Field-Oscillator Measurements

Two types of height/gain curve, using the signals from Start Point, are shown in Fig. 7(*a*). A definite ripple superimposed

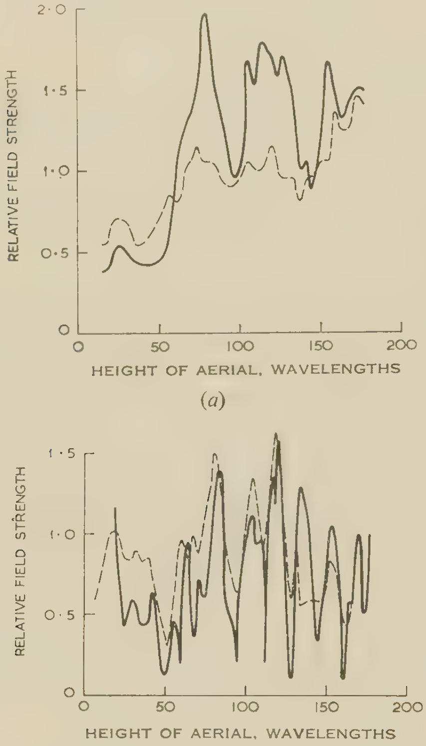


Fig. 7.—Variation of field strength with height (Ipswich).
24th April, 1958.
(*a*) Start Point signal. ——— 18.30 G.M.T.; ---- 19.30 G.M.T.
(*b*) Local transmitter signal. ——— Horn.; ---- Paraboloid.

on a steady rise of about 10 dB was present in both cases. The deep belt of trees (at a distance of 2·5 km) could produce sufficient diffraction loss to account for the 10 dB, and the ripple, when present, may have been due to interference between the direct and ground-reflected rays corresponding to the steady fraction of the signal.[6] Certainly, when a calibrating oscillator was used in a fixed position at a distance of 2·5 km, a pronounced standing-wave pattern was obtained, as shown in Fig. 7(*b*). The spacing between successive maxima altered with height but corresponded generally to a ground slope, in the foreground, of 1·3°. A survey confirmed that this was reasonable. Further, it will be seen that considerably deeper nulls in the pattern were obtained with horn receiving aerials than with paraboloids, since the beam width of the former was greater and therefore encompassed more of the ground-reflected component.

(4.3) Amplitude and Fading Measurements

As at Slough (Section 3.3), pen-recordings of mean amplitude in each channel were made whenever the equipment was running, and hourly samples of the fast fading were taken on film.

An estimate of the transmission loss for the period April–June is given in Table 2.

Table 2

MONTHLY MEAN TRANSMISSION LOSS

Month (1958)	Measured transmission loss (dB)	Derived basic path attenuation (dB)
April	164	228
May	159	223
June	159	223

Nominal gain of transmitting aerial = 36·5 dB.
Nominal gain of receiving aerial = 29 dB.

The figures in the right-hand column are calculated from the monthly measurements assuming 0 dB gain for both the transmitting and receiving aerials and also that a total aerial-medium coupling loss of 1·5 dB was present.

During the preceding March, while the equipment was being set up, some exceptionally low values of field were maintained for several days at a time, appearing to coincide with a dry spell of weather with cold northerly winds. It was difficult to estimate the magnitude at the time but the signals appeared to be about 5 dB lower than the usual minima (hourly mean values) subsequently observed. These weather conditions also occurred later and always tended to produce lower field strengths, possibly owing to the absence of super-refraction or to partial screening of the scattering volume by a high-level temperature inversion layer. Such weather was usually anticyclonic, and a tendency for low signals to occur sometimes during anticyclonic conditions has been observed[6] at this wavelength.

Fading rates again lay in the general region of 1–10 per second. According to Gordon's scattering theory, the horizontal angle through which power was mainly received should now have been 2·5°, about 1·5 times that at the range of Slough. Consequently, slightly faster fading might be expected,[4] but the variation in fading rate at each site was so great that it is doubtful whether this could have been satisfactorily resolved except by measurements made simultaneously at both stations. There is some evidence (Fig. 6) to indicate that the fastest fading is accompanied by a low correlation coefficient at a fixed aerial spacing, and therefore by a low correlation distance; consequently it is considered that the relationship shown in Fig. 5 for the Slough results may also be generally representative of the Ipswich data.

(4.4) Comparison with Vertical-Incidence Sounding at Slough

Slough was approximately at the mid-point of the Start Point–Ipswich path. An S-band vertical-incidence pulse sounder was being developed there at the time of the present experiments to observe discontinuities in the tropospheric refractive-index structure, and was occasionally in use. On six occasions it was possible to compare the Ipswich measurements with those made at vertical incidence at Slough. On only one of these could any striking effects be said to be common to both. A very unusual type of rhythmic fading with a period of about 4 min occurred at Ipswich at the same time as the height of reflection at Slough was alternating between a high and a low value at about the same rate. The upper reflection height was high enough to be within the common scattering volume on the Start Point to Ipswich path. If this connection is, in fact, real it provides direct evidence that unusual stratification conditions in the common volume affect, as might be expected, the nature of the received signal over a long-distance trajectory.

(5) MEASUREMENTS AT BLANDFORD (130 km)

The Blandford site was 107 m above sea level with a clear view towards the transmitter except for a 45 m hilltop at a distance

of 12 km. The angular distance over an unobstructed horizon would have been 16 mrad, but in fact was 19 mrad because of the intervening rise. The path was mostly over sea (transmitter–coast: 87 km; coast–Blandford: 43 km). Measurements were made for 3 days per week during September–December, 1958. The square pyramidal-horn receiving aerials were used, except for a short period when a fading-rate comparison between a horn and a paraboloid was carried out. As in the previous work, the horn apertures had sides 2λ long and the paraboloid was 13λ in diameter.

(5.1) Correlation between the Fast Fading in the Outputs of Two Spaced Receiving Aerials

(5.1.1) Vertical Aerial Spacing.

The lower aerial was at a fixed height of 1·5 m above the ground, at the base of the mast in this case. Some extreme results of correlation-coefficient/spacing runs are given in Fig. 8; the correlation coefficient fell substantially to zero at spacings varying from 10λ up to a maximum of over 100λ. The upper curve was, however, most unusual and was obtained during a spell of exceptionally steady signals. As at the other sites, a particular type of signal usually persisted for an hour or more, unless weather conditions were unstable; an unusual example of such instability will now be described.

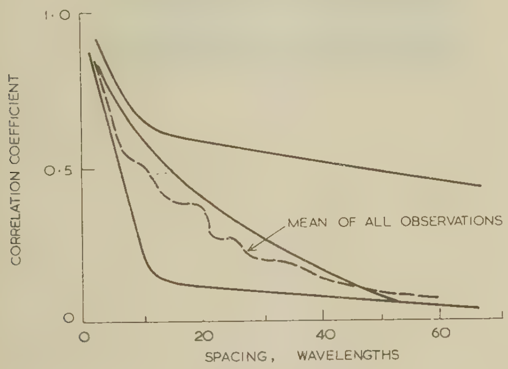


Fig. 8.—Variation of correlation coefficient with vertical spacing (Blandford).

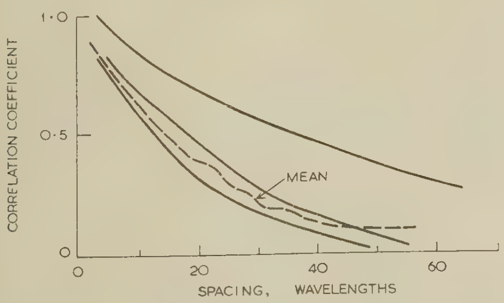


Fig. 9.—Variation of correlation coefficient with horizontal spacing (Blandford).

On 30th September (when the area was near the centre of a deep general weather depression) heavy rain squalls associated with the passage of a local depression swept across the valley in front of the receiving system. The correlation coefficient fell rapidly from near unity to a low value and rose again during the passage of a squall. On one occasion, r fell to zero, as can be seen from Fig. 10. The horns were adjacent, one above the other, during these observations, the spacing between centres being therefore only 2λ. The spread of incident energy in the vertical plane must have been at least 15°, and no appreciable change in the mean level of the signal received on each horn was observed; this is not inconsistent if it is recalled that the beam width of each horn was 32°. (Presumably a similar spread of energy would have occurred in the horizontal plane, but these unusual weather conditions did not persist sufficiently long for this to be confirmed experimentally.) The lower diagram in Fig. 10, relating to data obtained later in the day, shows that a fall in correlation coefficient was accompanied by a rise in the fading rate, as might be expected. It is evident that, given unusual weather conditions, scatter signals can be substantially uncorrelated at aerial spacings as small as 2λ.

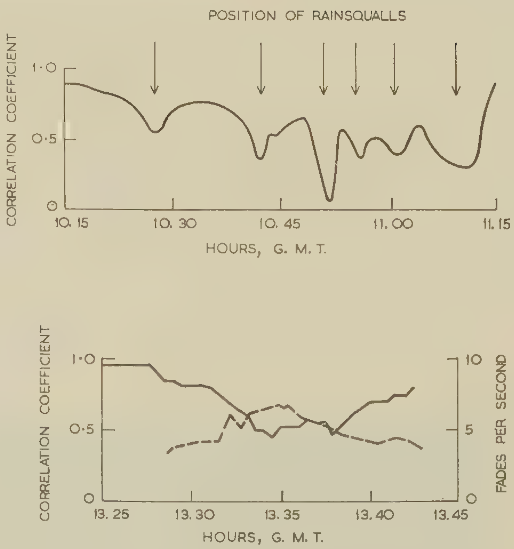


Fig. 10.—Effect of local rainsqualls (Blandford). 30th September, 1958.

(5.1.2) Horizontal Aerial Spacing.

Some typical results with varying horizontal spacings are given in Fig. 9. The signals were usually substantially uncorrelated at spacings varying from 20λ to perhaps 150λ. There was thus again a tendency for larger spacings to be required horizontally than vertically to achieve conditions giving substantially uncorrelated amplitude fluctuations from the two aerials, although this was not as marked as at the longer-range sites.

(5.2) Height-Gain Measurements and Variation of Field-Strength with Horizontal Position

Two extreme cases of the variation of field strength with height are shown in Fig. 11. It was found at this site that persistently steady signals were often obtained (see Section 5.3). Such signals produced a standing-wave pattern in the vertical plane as shown in Fig. 11, attributable to interference between a direct ray and one reflected from the sloping ground in front of the aerial system. The spacing between successive maxima was 15λ around heights of 25λ, falling to 10λ at heights in the vicinity of 80λ. Those results could be brought about by reflections from ground having a slope of 2°, 30 m in front of the aerial system, and gradually rising to 2·9° at a distance of 60 m. These figures check quite closely with the ground-slope variation deduced from the map contours. No mean height gain over this height range was found, which is not surprising since the nearest obstruction was at a distance of 12 km.

Evidence of standing waves was also found along a horizontal line broadside to the great-circle direction. As with the measurements in the vertical plane, one aerial was always stationary

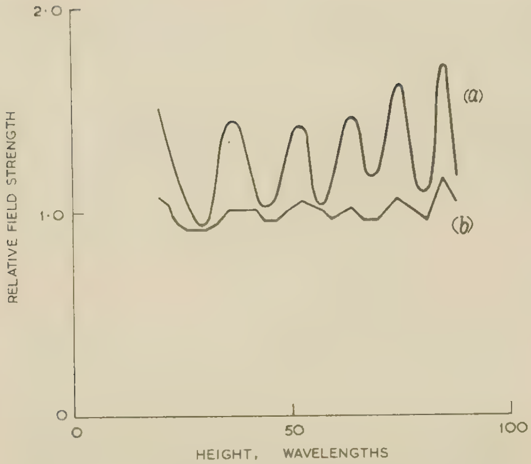


Fig. 11.—Variation of field with height (Blandford).
(*a*) 13th October, 1958.
(*b*) 15th October, 1958.

(Section 3.2) and was used to enable temporal variations due to slow fading to be allowed for. Some extreme results are given in Fig. 12 for measurements made at a height of 19·5m. Also shown is a curve obtained when a calibrating transmitter was used over a line-of-sight path 11·3km in length.

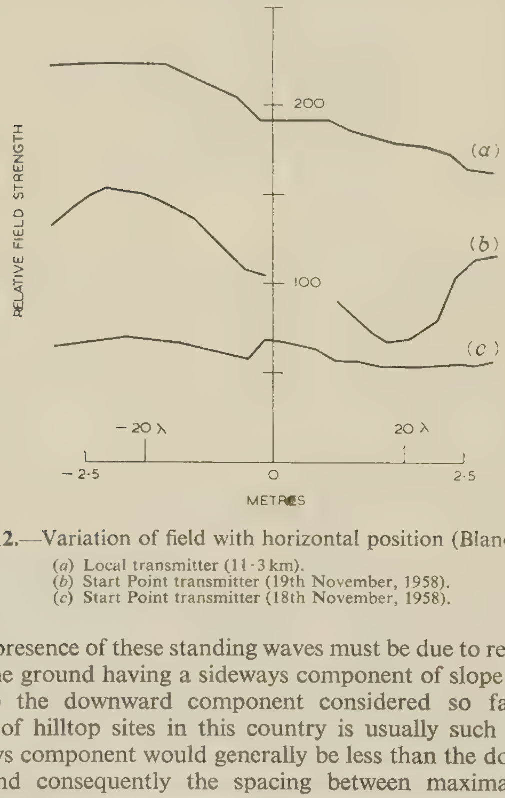


Fig. 12.—Variation of field with horizontal position (Blandford).
(*a*) Local transmitter (11·3 km).
(*b*) Start Point transmitter (19th November, 1958).
(*c*) Start Point transmitter (18th November, 1958).

The presence of these standing waves must be due to reflections from the ground having a sideways component of slope in addition to the downward component considered so far. The nature of hilltop sites in this country is usually such that the sideways component would generally be less than the downward one, and consequently the spacing between maxima in the horizontal plane might generally be expected to be greater than that in the vertical plane at a given site. This certainly appears to be so at Blandford, where the spacing between maxima in the vertical plane was 10–15λ, whereas in the horizontal plane it was 40λ. The variation between successive days in the Start Point curves (Fig. 12) suggests the presence of a strong coherent component in the signal on 19th November, 1958, and a weak one on 18th November, 1958.

(5.3) Amplitude and Fading Measurements

Together with the pen-recordings of field strength and th fading samples taken as at other sites, some recordings of th normalized variance of the fast fluctuations of the signal ampl

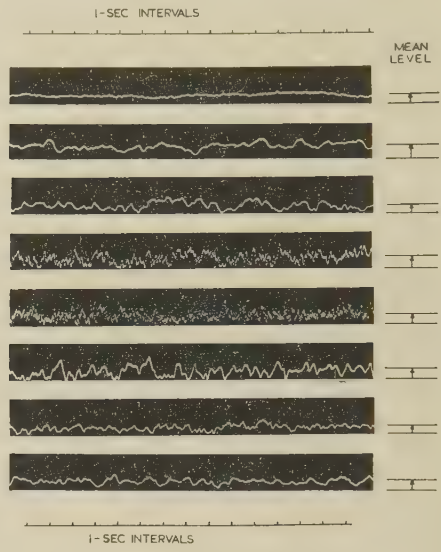


Fig. 13.—Samples of fading signal (Blandford).
The voltage scale is linear.

tude were made by the method indicated in the next paragrap This was done because it was found that at times the signal wa quite steady whereas at other times it suffered the more usu near-Rayleigh[6] type of fading. This is illustrated in Fig. 1 which shows how the type of fading could change steadi throughout the day (there is a 3h interval between successi pictures), in this case without much alteration in the mean lev of the signal.

The normalized variance was defined as the mean-squa fluctuation about the mean divided by the mean-square sign level. For fading following an amplitude distribution based c the vector sum of a random and a steady component,[11] th greatest value of normalized variance which can be obtained 22%, which occurs when the steady component is zero, a Rayleig distribution then remaining. Pen-recordings of the square the amplitude fluctuations about the mean and of the squa of the total signal amplitude were made over 5min perio using the squaring circuits of the correlator with and without t d.c. rejector respectively.[8] Some results are shown in Fig. 1 where it will be seen that the variance rises to the theoretic maximum corresponding to a Rayleigh distribution of amp tudes. The minimum value ever recorded was 0·01. Fadi rate is also plotted on this diagram, and the similarity between t general trends of fading rate and variance in this case is evide In detail, whenever very high fading rates occurred the normaliz variance was also high, although the occurrence of a hi variance was not necessarily associated with a high fading ra at a particular time, but tended to be so on the average.

A short fading test was performed in which signals we received simultaneously on the paraboloid (5° beam wid between 3dB points) and the horn (32° beam width betwe 3dB points), which were adjacent one above the other. T results were recorded side by side on a single film strip, a samp of which is shown in Fig. 15. Additional small faster ripp are evident on the horn output, so some energy was eviden being received in angular spreads exceeding 5°. This at fi

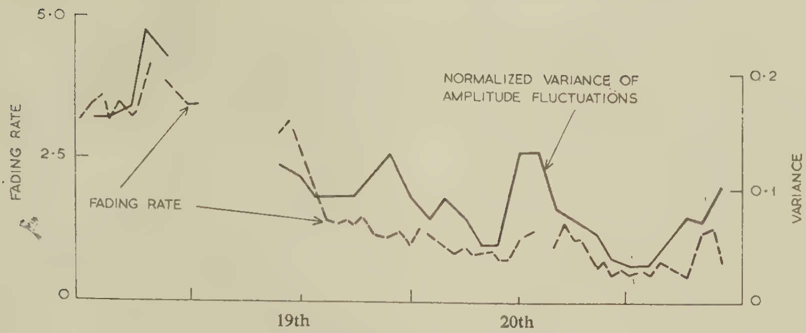


Fig. 14.—Variance and fading rate (Blandford).
18th–21st November, 1958.

sight appears rather large and at variance with the estimates of angular spread already made by correlation methods, but the ripples were in fact so small that they did not seriously affect the value of correlation coefficient existing between the outputs of the two aerials.

A comparison between the variations in the hourly mean values of field strength occurring at Blandford and those occurring simultaneously at Wembley* (279 km) was made over the whole 3-month period. There is little evidence of correlation at this separation (149 km). (In a similar comparison[7] between Wembley and Slough at a separation of 21 km a substantial correlation was found.) There were fewer signal enhancements[6] at Blandford than at Wembley and they occurred at random times, whereas those at Wembley had a tendency to occur during the hours around midnight. The total decibel spread in the slow fading was also lower at Blandford, as can be seen from Table 3;

Table 3

HOURLY MEAN TRANSMISSION LOSS AT WEMBLEY AND BLANDFORD

(3 month period)

Site	Measured transmission loss (dB)				Derived basic path attenuation (dB)			
	Max.	Min.	Mean	Spread	Max.	Min.	Mean	Spread
Blandford	154	134	145	20	207	187	198	20
Wembley	163	130	148	33	232·5	199·5	217·5	33

Nominal gain of transmitting aerial = 36·5 dB.
Nominal gain of receiving aerial at Blandford = 17 dB.
Nominal gain of receiving aerial at Wembley = 36·5 dB.

this was owing primarily to the fact that the individual enhancements at Wembley were bigger than those at Blandford. However, in making any comparison between measurements at Blandford and at Wembley it must be remembered that the former relate to a path which is substantially over sea and the latter to one which is mainly over land.

A coupling-loss figure of 3·5 dB[6] was assumed in deriving the basic path attenuation for the Wembley results, and a figure of 0·5 dB was estimated for Blandford from the angular spreading of energy found in the present paper. However, since the spread was variable the gain degradation was itself variable, depending on propagation conditions.

The mean difference between the transmission losses at the two distances was 19·5 dB. Assuming that the scattered power tends to be inversely proportional to the fourth power of the scattering angle,[5] had the Blandford path been an unobstructed

* Wembley recordings were provided by the G.E.C. Research Laboratories.

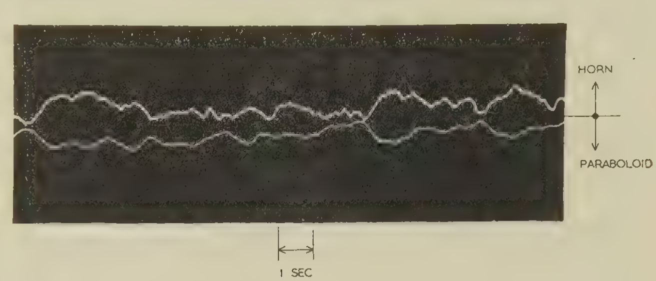


Fig. 15.—Fading comparison between horn and paraboloid. (Blandford).

one the scattered power would have been increased by nearly 3 dB, so the mean difference would then have been 22·5 dB. Assuming further that the attenuation of power with distance, d, varies as $1/d^n$, this then gives $n = 6 \cdot 8$, in close agreement with $n = 7$ found by Gordon[5] for large-scale scatterers. A value of 7 for n means that the path attenuation is increased by 21 dB each time the distance is doubled; consequently, an additional path attenuation of 21 dB should be introduced by removing the receiver from Blandford to Slough, and a further 13 dB by removing from Slough to Ipswich; but the corresponding measured values (taken from Tables 1 and 2 and empirically modified for seasonal variation in the Wembley 'control' measurements) turn out to be only 18 and 10 dB respectively. Consequently n appears to be nearer 6 than 7 for these results, which were not, it must be emphasized, obtained simultaneously at all sites.

The monthly mean figures of transmission loss (measured using the 36·5 dB transmitting aerial and the 17 dB receiving horn) for the months of September, October and November, 1958, were respectively 146, 145 and 144 dB at Blandford.

(5.4) Auto-Correlation and Cross-Correlation Functions of the Amplitude of the Fading Signals

The signals from the two horns (spaced horizontally at 4·9 m) were recorded separately on the channels of a twin-channel tape-recorder by the method described in Section 2.2.3. A number of recordings corresponding to different conditions were subsequently selected and turned into auto-correlation and cross-correlation functions. Examples of the latter are shown in Fig. 16. A maximum was usually present in the function, and a standard test showed it to be significant.

The presence of a maximum in the cross-correlogram, which is displaced from zero time, suggests the movement of a diffraction pattern along the line of the aerials, i.e. across the propagation path. This movement almost certainly arises from a cross-path drift due to wind in the scattering volume as in the iono-

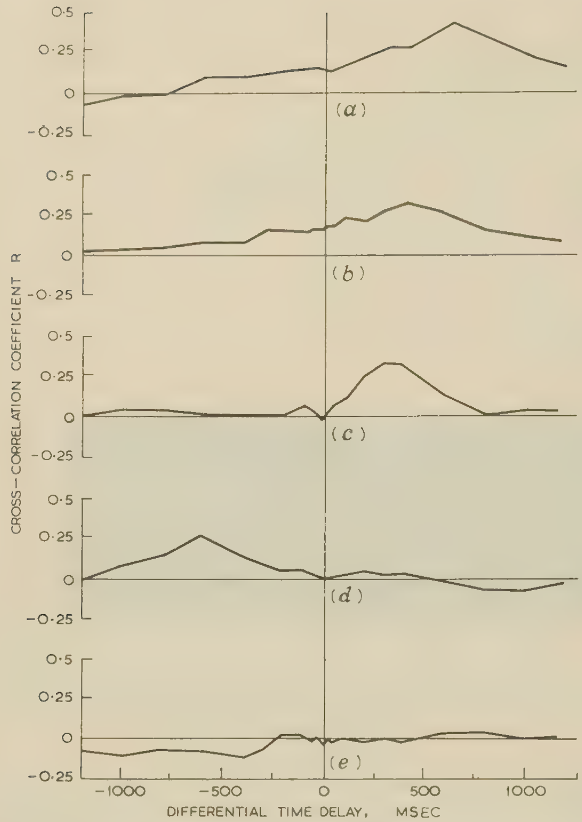


Fig. 16.—Variation of cross-correlation with differential time delay (cross-correlograms).

(*a*) 7th November, 1958; 06.00 G.M.T.; fades per second: 1·3.
(*b*) 7th November, 1958; 04.00 G.M.T.; fades per second: 2·0.
(*c*) 13th November, 1958; 02.00 G.M.T.; fades per second: 7.2.
(*d*) 19th November, 1958; 12.00 G.M.T.; fades per second: 3.1.
(*e*) 12th November, 1958; 08.00 G.M.T.; fades per second: 11·8.

spheric case. The determination of the wind speed and direction from a cross-correlogram is a complicated matter[12] and requires three receiving aerials together with a knowledge of the corresponding auto-correlation function. However, assuming for present purposes that the scattering was taking place near the mid-path so that the diffraction pattern was moving across the receiving aerials at twice the rate of drift, a transverse velocity of, for example, 1 m/s should correspond very roughly to a time delay of 2·5 sec. This could only be confirmed approximately since the available meteorological data were insufficiently localized; however, there was little doubt that when the transverse velocity was reversed the sign of the delay required for maximum cross-correlation also reversed.

Very few auto-correlation functions have so far been examined, but there is evidence of a tendency for a greater initial negative slope in these functions to be accompanied by smaller values of optimum delay in the corresponding cross-correlation functions. This is to be expected since a greater slope implies a higher fading rate, and a smaller delay, a higher wind velocity, a connection between high fading rates and high wind velocities having been mentioned in Section 3.3.

(6) CONSIDERATION OF RESULTS

The correlation-coefficient/spacing results were variable and did not usually appear in a form which suggested a definite law, such as a Gaussian-shaped angular distribution of incident energy; the picture is rather one of a 'lumpy' medium in which the positions of major scattering areas or other discontinuities were changing appreciably during the length of time required to perform a correlation/spacing run. Generally speaking, however, although the results were variable the correlation distances appeared (at any rate at Slough) to be greater in the horizontal than in the vertical direction, implying that the angular spread of the incident energy was greater in the vertical than in the horizontal plane. The Booker–Gordon scatter theory, however, indicates the reverse. An explanation is possible if the presence of a number of tilted reflecting surfaces is assumed.

From considerations of path geometry, the value of tilt, t, of such a layer about the cross-path horizontal, required to produce a specular reflection in the vertical plane through the great-circle path at an angle of elevation, δ, at the receiver, can be shown to be approximately $\delta^2/2(\delta + \theta_c)$ radian, where θ_c is the mid-path scattering angle and all angles are small; the surface is assumed to be on or just above the transmitter horizon plane, and the maximum permissible angle of elevation at the receiver is determined by the cut-off of the polar diagram of the receiving aerial. Alternatively, the angle of elevation produced by the tilt is given by $t + \sqrt{(t^2 + 2\theta_c t)}$. Lateral deviation of the incident ray can be produced by a cross-path tilt but the deviation produced by such a tilt when across the mid-path is only $t\theta_c/2$ radian, which is very much smaller than δ. Consequently, the frequency of occurrence of substantially deviated specular reflections is likely to be greater in the vertical than in the horizontal plane. If such specular reflections do in fact occur, they should usually appear to originate close to the great-circle direction, according to the foregoing argument; this effect was observed very clearly with ionospheric sporadic-E layer reflections at oblique incidence.[13]

As a numerical example, at Slough the tilt required to produce a value of δ equal to 2·5° is about 1°. This magnitude of tilt can produce a horizontal deviation of only 1′ of arc, compared with 2·5° in the vertical plane. Specular reflections, whether of this type or brought about by, say, super-refraction phenomena are necessary to account for the ripples which have been observed in the height-gain curves and also for the steady background signal which produces an alteration of slope in the Rayleigh paper plots of the instantaneous signal amplitude distribution described elsewhere.[6] An alternative explanation leading to smaller correlation distances in the vertical than in the horizontal has been suggested by Staras,[14] who assumes anisotropic scattering from turbulence having greater intensity in the vertical than in the horizontal plane. However, it is not immediately obvious that such a statistical scatter theory can lead to a specular or steady signal component similar to that often observed.

The presence of some kind of layer formation is feasible since stratification has been observed at Slough at vertical incidence; further, the possibility—perhaps a surprising one—of discontinuous surfaces existing at least for a time during windy conditions appears to be shown by the fact that the cross-correlation coefficient between the outputs from a pair of spaced aerials is often significantly greater than zero at the appropriate value of differential time delay.

There is, however, another effect which would tend to produce greater *apparent* deviations in the vertical than in the horizontal plane, and is therefore worth mentioning although it does not turn out to be important. If the angular distribution of amplitudes in the downcoming scatter is assumed to be Gaussian, it can be shown that reflections from a foreground having a downward slope cause a ripple to occur in the vertical-plane r distribution that would be obtained if the ground were horizontal; if the reflection coefficient at the ground is unity the minima of the ripples become nulls. The effect of the first minimum introduced by ground of moderate tilt is therefore to bring

towards zero at smaller values of spacing than would have occurred with horizontal ground, which would be interpreted in the present paper as implying a larger spread of downcoming scatter energy (or greater deviations) in the vertical plane. However, this effect cannot be invoked to explain the Slough results since no appreciable ground slope existed there: at Blandford, where a sloping foreground did exist, the mean of 40 vertical spacing runs (plotted in Fig. 8) shows definite ripples which are almost certainly due to the presence of ground reflections in the manner described, but they are too small to be of much consequence. It is therefore inferred that the effect has not been an important one.

It is interesting to speculate on a model for a statistical analysis based on a finite number of relatively strong rays or cones of rays emanating from discontinuities which are moving randomly with respect to each other while having a uniform drift. Such strong rays may occasionally be formed from slowly undulating air surfaces which at times are so shaped as to form a coherent reflection, leading (on the average) to localized tilted reflecting surfaces giving greater deviations in the vertical than in the horizontal plane according to the mechanism described in the previous paragraph; the common volume is enormous and the chances of a number of such patches being present simultaneously is probably large. If, however, the number remains fairly small there is a good chance that one will always be appreciably greater than the rest and produce the background signal referred to earlier in this Section. In calm weather the length of time for which the discontinuities persist might be large, whereas in conditions of great turbulence, such as must have occurred in the rain squalls described in Section 5.1.1, rapidly changing conditions giving many paths may occur, approximating to a mechanism of statistical scattering.

No consistent relationship between the signal amplitude, fading rate and correlation distance has been observed, although there was a tendency for the highest fading rates to be accompanied by low correlation distances. On the shortest path, which was largely over sea, substantial variations in the normalized variance of the signal-level fluctuations occurred, the signal being steady at times and at others showing fast Rayleigh-type fading. The slower variations in variance and fading rate were highly correlated, faster fading tending to give higher variance.

Despite the great variability in the effects observed, one phenomenon occurred quite consistently. At the two sites (Blandford and Ipswich) where the receiving system overlooked a sloping foreground, definite indications of standing waves due to ground reflections were present. In fact, with the steady signal from a distant oscillator over a line-of-sight path, the standing waves were deep and showed that the ground reflection coefficient was as high as 0·9 at times. The ground contours were typical of the gently rolling foregrounds associated with sites in southern England, and on such foregrounds the chances of finding a patch of ground oriented to produce a strong reflection in the direction of the aerial system may be great and should be borne in mind. Owing to the angular spread of energy in the downcoming wave these reflections will not normally be serious on long scatter paths; but over short paths, where substantially single-ray transmission may occur, interference from the reflected component may lead to considerable reduction in the resultant signal if the aerial height is incorrectly chosen.

(7) CONCLUSIONS

Measurements of amplitude, fading characteristics, correlation distance and height gain have been made successively at three distances (258, 398 and 130 km) from a continuous-wave transmitter. The time spent at each site was 2–3 days per week over a period of about three months, sufficient to enable a fair idea of the short-period variations to be obtained. However, the overall changes in the quantities being measured turned out to be great, and, owing to the relatively short time spent at each site, a comprehensive statistical analysis of the various quantities has not always been attempted but extreme values have been indicated. These may be useful to system designers since, although they are extremes, they have usually persisted for considerable lengths of time and would therefore have to be taken into account.

Although the results were often difficult to interpret it is considered that, for the medium range (258 km), the angle containing most of the incident fluctuating energy in the vertical plane was not usually less than 1° and not usually greater than 3°; in the horizontal plane the corresponding figures were 0·5° and 1·5°. The latter is in reasonable agreement with Booker–Gordon scatter theory, but the range of vertical angles found was much greater, as might be expected from a layer-type theory. The aerial separations for which the signal outputs were substantially uncorrelated were 10–30λ vertically and 20–60λ horizontally. At the shortest range the separations required appeared to be somewhat larger, perhaps 150λ in the horizontal plane, again in general agreement with the scatter theory. Consequently a spacing of 150λ should be adequate, at least for the present system, at most ranges for two horizontally spaced aerials operated in diversity, with a tendency for smaller spacings to be required in the vertical direction and at the greater ranges.

Measurements have shown that a consistent height gain was present at the two sites where standard diffraction theory based on the presence of fairly local obstructions indicated that it might have been expected. Superimposed on this general increase of field strength with height were ripples due to reflections from the ground, and the reasons for the presence of these have been discussed.

It appears generally from the results that no simple model for the troposphere is likely to account for the type of variations observed, and the implications of this have been discussed briefly. It must be concluded that there is a need to examine the fine structure of the scattering medium if further substantial information on the nature of the scatter mechanism is required. This might be done by using equipment which scans the incident energy with a very narrow beam[14] or by pulse-sounding equipment.

(8) ACKNOWLEDGMENTS

The transmitter at Start Point was constructed and operated by the Research Laboratories of the General Electric Co. Ltd. under contract to the Admiralty and Ministry of Supply, and the co-operation of these bodies is gratefully acknowledged. The measurements were carried out largely by Mr. J. Bell and Mr. I. D. Clough, and the work forms part of the programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(9) REFERENCES

(1) Booker, H. G., Ratcliffe, J. A., and Shinn, D. H.: 'Diffraction from an Irregular Screen with Applications to Ionosphere Problems', *Philosophical Transactions of the Royal Society*, 1950, **242**, p. 579.

(2) Megaw, E. C. S.: 'Fundamental Radio Scatter Propagation Theory', *Proceedings I.E.E.*, Monograph No. 236 R, May, 1957 (**104** C, p. 441).

(3) Bain, W. C.: 'The Angular Distribution of Energy Received by Ionospheric Forward Scattering at Very High Frequencies', *ibid.*, Paper No. 2487 R, January, 1958 (**105** B, Suppl. 8, p. 53).

(4) NORTON, K. A., RICE, P. L., JANES, H. B., and BARSIS, A. P.: 'The Rate of Fading in Propagation Through a Turbulent Atmosphere', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 1341.

(5) GORDON, W. E.: 'Radio Scattering in the Troposphere', *ibid.*, 1955, **43**, p. 23.

(6) ANGELL, B. C., FOOT, J. B. L., LUCAS, W. J., and THOMPSON, G. T.: 'Propagation Measurements at 3 480 Mc/s over a 173-mile Path', *Proceedings I.E.E.*, Paper No. 2510 R, January, 1958 (**105** B, Suppl. 8, p. 128).

(7) GEIGER, G. V., LA FRENAIS, N. D., and LUCAS, W. J.: 'Propagation Measurements at 3 480 and 9 640 Mc/s beyond the Radio Horizon', *ibid.*, Paper No. 3319 E, November, 1960 (**107** B, p. 531).

(8) BELL, J.: 'Correlation between Fading Signals', *Electronic Technology*, January, 1960, **37**, p. 36.

(9) NORTON, K. A., RICE, P. L., and VOGLER, L. E.: 'The Use of Angular Distance in Estimating Transmission Loss and Fading Range for Propagation through a Turbulent Atmosphere over Irregular Terrain', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 1488.

(10) KITCHEN, F. A., JOY, W. R. R., and RICHARDS, E. G.: 'Influence of the Semi-Permanent Low-Level Ocean Duct on Centimetre Wave Scatter Propagation beyond the Horizon', *Nature*, 1958, **182**, p. 385.

(11) MCNICOL, R. W. E.: 'The Fading of Radio Waves of Medium and High Frequencies', *Proceedings I.E.E.*, Paper No. 891 R, November, 1949 (**96**, Part III, p. 517).

(12) BRIGGS, B. H., PHILLIPS, G. J., and SHINN, D. H.: 'The Analysis of Observations on Spaced Receivers of the Fading of Radio Signals'. *Proceedings of the Physical Society*, **63** B, 1950, p. 106.

(13) MEADOWS, R. W.: 'The Direction and Amplitude of Reflections from Meteor Trails and Sporadic-E Ionization on a 1 740 km North-South Path at Very High Frequencies', *Proceedings I.E.E.*, Paper No. 2537 R, January, 1958 (**105** B, Suppl. 8, p. 56).

(14) STARAS, H.: 'Forward Scattering of Radio Waves by Anisotropic Turbulence', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 1374.

(15) WATERMAN, A. T., JNR.: 'A Rapid Beam-Swinging Experiment in Trans-horizon Propagation', *Institute of Radio Engineers Transactions on Antennas and Propagation*, October, 1958, **AP-6**, p. 338.

DISCUSSION ON 'TELEVISION BAND COMPRESSION BY CONTOUR INTERPOLATION'*

Dr. D. A. Bell (*communicated*): The system proposed has the practical advantage that the basic operation, namely deleting from the transmission $n-1$ out of every n frames, is systematic and therefore involves no problem of synchronizing information; and its storage requirements are fully determined. But its defect is that it deletes part of the information capacity of the system, whether or not the full capacity is occupied by the original signal at any instant. This is in contrast to systems such as velocity modulation (which was suggested independently by Cherry and Gouriet[A] and by the writer[B]) or run-length coding.[C] A systematic reduction risks losing valid information, whereas a reduction based on the absence of valid information risks running out of storage on occasion.

The authors claim that their scheme, which utilizes a combination of line-to-line correlation and frame-to-frame correlation, can be combined with means of using within-line correlation to give the product of the compression ratios of the two. Is this quite fair? Does it assume that the two correlations are independent?

A very valuable feature of the authors' scheme is that they have defined 'contour' in terms of rate of change of intensity, and thus avoided quantizing picture amplitudes. It is impossible to judge picture quality from the reproductions in the *Proceedings*, which necessarily have a process-engraver's screen superimposed, but they seem to be free from the spurious contours which arise from quantization.

REFERENCES

(A) CHERRY, E. C., and GOURIET, G. C.: 'Some Possibilities for the Compression of TV Signals by Re-coding' in 'Communication Theory' (Ed. Willis Jackson) (Butterworths, 1953).

(B) BELL, D. A.: 'Information Theory and its Engineering Applications' (Pitman, 1st ed., 1953).

(C) SCHREIBER, W. F., and KNAPP, C. F.: 'TV Bandwidth Reduction by Digital Coding', *I.R.E. Convention Record*, 1958, **6**, Part 4, p. 88.

* GABOR, D., and HILL, P. C. J., Paper No. 3507 E, May, 1961 (see **108** B, p. 303).

Prof. D. Gabor and **Dr. P. C. J. Hill** (*in reply*): Dr. Bell queries the possibility of combining compression methods which make use of within-line correlations with ours, which makes use of line-to-line and frame-to-frame correlations, with a resulting gain which is simply the product of the gains. This is an interesting point, but the answer is clearly in the affirmative. 'Within-line' methods make no use whatever of line-to-line and frame-to-frame correlations; the coding starts afresh in every line, regardless of what happened in the previous one; hence our method taps a new source of savings. This does not mean, however, that 'the correlations are independent'; it means only that we have analysed a three-dimensional correlation in terms of three components. We can obtain a general idea of this three-dimensional correlation as follows. There is an abrupt transition in one line. It is then highly probable that there is a similar abrupt transition somewhere near in the preceding and the following lines (because the abrupt transition probably represents a more or less vertical contour), and also in the corresponding lines of the previous and the following frames (because the contour belongs very probably to a permanent object). This suggests that there may exist methods more economical than the piecemeal extraction of correlations in the three dimensions by separate apparatus, but we have made it clear that, whatever the scientific interest of the subject may be, its practical application appears highly doubtful.

621.396.97: 621.391.812.63

The Institution of Electrical Engineers
Paper No. 3506 E
Feb. 1961

AN INVESTIGATION OF THE USEFULNESS OF BACK-SCATTER SOUNDING IN THE OPERATION OF H.F. BROADCAST SERVICES

By E. D. R. SHEARMAN, B.Sc.(Eng.), Associate Member.

(*The paper was first received* 6*th July, and in revised form* 28*th November*, 1960. *It was published in February*, 1961, *and was read before the* ELECTRONICS AND COMMUNICATIONS SECTION 6*th March*, 1961.)

SUMMARY

In h.f. broadcasting, frequencies and aerials are chosen to give the optimum signal strength to listeners within the desired service area. The varying nature of the ionospheric transmission medium makes it desirable to measure the distribution of energy actually achieved over that area.

In the experiments described, pulse echo measurements of the back-scattered energy from ground irregularities in the service area were made, the echoes being observed on a range/amplitude display. In tests with two aerial systems, the echo patterns clearly showed the differing energy distribution. The patterns agreed well with those calculated for the two aerial systems from parabolic ionospheric-layer theory and the radar equation.

Further tests at night under varying ionospheric conditions showed that the propagation modes to 5 000 km range were correctly evaluated by back-scatter sounding at times when ionospheric predictions, corrected by local vertical-incidence soundings, gave erroneous results.

LIST OF PRINCIPAL SYMBOLS

D = Ground range for a single ionospherically reflected hop.
D_1 = Part ground range under section of ray path in ionosphere.
D_2 = Part ground range under free-space section of ray path.
R = Radius of earth.
h_0 = Height of bottom of ionosphere (assumed parabolic) above ground.
i_0 = Angle of incidence of ray on bottom of layer.
h_m = Height of maximum ionization of layer.
y_m = Semi-thickness of layer (assumed parabolic) $= h_m - h_0$.
$x = \dfrac{\text{Transmitted frequency}}{\text{Critical frequency of layer}}$.
Δ = Angle of elevation of ray at ground.
p' = Equivalent free-space path of ray.
p_1' = Equivalent free-space path of part of ray in ionosphere.
p_2 = Path of ray below layer.
r = Range of radar target.
P_T = Transmitter radiated power.
G_T, G_R = Power gains of transmitting and receiving aerials, respectively, relative to an isotropic radiator.
λ = Wavelength.
A = Echoing area of target.
c = Velocity of light.
δt = Duration of radiated pulse.
θ_B = Half-power beam width of aerial for common transmit-receive.

The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

(1) INTRODUCTION

The paper gives an account of two series of experiments carried out in February and March, 1957, to evaluate the possible usefulness of back-scatter sounding in the operation of the oversea broadcasting services of the British Broadcasting Corporation. The techniques employed and the results obtained should also be of interest in the study of other h.f. broadcast services and point-to-point communication networks.

Back-scatter sounding, an application of radar technique to the study of h.f. propagation, has been investigated by a number of workers in recent years[1,2,3] and the literature of the subject has been surveyed in a recent paper.[4]

In back-scatter sounding, as in radar, pulses of waves are radiated from a directional aerial and the ranges of targets are determined from the time delay of the returning echoes. In contrast to radar, however, the primary purpose is not to determine the location of the targets, which in scatter sounding are the irregularities present over the whole surface of the earth, but to find which parts of the earth are illuminated by ionospheric reflection of the transmitted radio waves. These illuminated zones of the earth have an azimuthal extent governed by the beam width of the transmitting aerial, and lie beyond the skip distance determined by the transmitted frequency and the prevailing ionization and height of the ionospheric layer.

Up to the present time, emphasis has been placed on the use of scatter sounding for the measurement of skip distance at a particular frequency, and hence, by observations at a number of frequencies, for the determination of the m.u.f. for a particular distance. Such measurements yield only limiting conditions, the minimum distance or maximum frequency for communication. It is now considered that the technique can also supply information about the distribution of energy over the illuminated zone, so that the field strength set up at various parts of the service area can be deduced.

In broadcasting services, frequencies and aerials are chosen to give the highest signal strength to listeners within a specified area. The measures taken to achieve this are necessarily determined by the present state of knowledge about the complex process of ionospheric propagation and by the accuracy of prediction of layer ionization and height. Measurements of the distribution of energy over the earth actually achieved are thus desirable. These can be provided by listeners' reports, but this method introduces delay and limits the observations to the specific localities of the observing stations.

The possibility of using back-scatter echo-patterns observed at the transmitting station to determine instantaneously the distribution of energy over the required service area thus has considerable attractions. It should be possible to observe immediately the changes in this energy distribution as the frequency of transmission and the vertical radiation pattern of the transmitting aerial are varied, and thus to arrive at the optimum conditions for broadcast coverage. Such observations on days of varying ionospheric conditions should enable the

measures necessary to allow for day-to-day variations to be found.

Accordingly, the objects of the experiments described in the paper were to find the location of the illuminated zones by back-scatter observations, and to study the extent to which the range/amplitude echo patterns gave a guide to the variation of field strength along the great circle through the transmitting-aerial beam. The effect on the observed patterns of different aerials was investigated, and the pattern shapes were compared with those calculated for the ionization and heights of the ionospheric layers as measured simultaneously by vertical-incidence techniques. These comparisons gave close correlation and suggested that scatter sounding was a good indicator of actual propagation conditions and service-area coverage.

In the second series of tests the co-operation of distant observers was obtained to report on signal strength and thus to provide a check of the field strength at the distant points. A frequency was chosen which, it was hoped, would give an unilluminated gap between the areas illuminated by 1-hop and by 2-hop rays, this gap embracing some of the observing localities. The existence of such gaps, which is not predicted if the usual 'two control-point' propagation assumptions are made, has been indicated by back-scatter and other observations at Slough and is in accord with a recently developed technique for the computation of multi-hop propagation.[5]

Unfortunately, ionospheric conditions during the tests were not suitable for the formation of a gap. The experiment was nevertheless useful as it provided the opportunity for comparing the day-to-day changes in the back-scatter echo-patterns with the corresponding observed changes in one-way propagation to the receiving stations, and with the propagation modes calculated by the technique already mentioned.

(2) EXPERIMENTAL TECHNIQUE

In both the February and the March tests, quarter-hour transmissions, beamed in a southerly direction, were made from the B.B.C. short-wave transmitting station at Rampisham, Dorset, with a pulse power of 50 kW. The pulse duration was 1 millisec and the pulse repetition frequency 25 pulses/sec, locked to the 50 c/s supply mains.

The back-scatter echo-patterns were observed at Slough and, for comparison, transmissions were made from Slough immediately after the Rampisham schedule. A pulse power of 100 kW with a pulse duration of 120 microsec and a repetition frequency of 25 or $12\frac{1}{2}$ pulses/sec was employed at Slough. The equipment used at Slough has been described elsewhere.[6]

For transmission from Rampisham and reception at Slough, the effective aerial directivity was the product of the Slough and Rampisham directivities (neglecting small differences between departure and arrival angles due to the geographical separation of the two stations). For transmission and reception at Slough, on the other hand, the effective aerial directivity was the square of the Slough directivity. Comparison of the back-scatter patterns under the two conditions could thus be used to study the effect of aerial directivity in the vertical plane on the back scatter, and to study the possibility of measuring actual aerial performance by the technique.

The longer pulse duration of the Rampisham transmission, 1 millisec compared with 120 microsec for Slough, was chosen to make use of the higher mean power available with the B.B.C. transmitters. Since the echo power received by back-scattering from the ground should increase, for short pulses, in proportion to the pulse width, it was expected that the 1 millisec pulses would give stronger echoes and thus enable returns to be detected from areas of lower field strength than for the Slough equipment.

With equal aerial gains an increase of $(50/100) \times (1\,000/120)$ $\simeq 4$ times in power or 2 times in amplitude might be expected.

(2.1) Test on 11th–15th February, 1957

The schedule of transmissions and the parameters of the transmitting and receiving equipment in this first test are given in Table 1, and the vertical radiation patterns of the Rampisham and Slough aerials for the operating frequency of 21·6 Mc/s are plotted in Fig. 1(*a*). The directions of shoot in azimuth of the

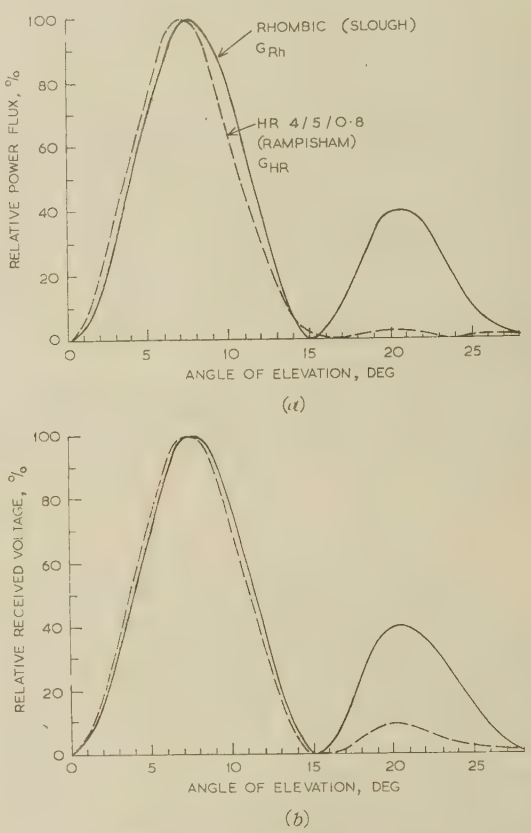


Fig. 1.—Aerial characteristics.

(*a*) Vertical-plane radiation pattern of Slough and Rampisham aerials (February 1957, test).
(*b*) Combined aerial characteristics for echo reception.
– – – – Transmission at Rampisham, reception at Slough ($\sqrt{G_{Rh}}\sqrt{G_{HR}}$).
——— Transmission and reception at Slough ($\sqrt{G_{Rh}}\sqrt{G_{Rh}}$).

two aerials differed by 7°, which was within the beam width for either aerial.

Synchronization between the transmitted pulses at Rampisham and the receiver time-base at Slough was achieved by using the 50 c/s supply mains as a trigger source at each station, a divider stage being inserted to give 25 c/s. This method proved satisfactory, the back-scatter echoes appearing in a stationary position on the receiver time-base without any detectable short-term or long-term variation. It may be mentioned that the mains-lock technique has been abandoned for oblique-incidence pulse observations of forward propagation in the United Kingdom, except for short-distance work, because of phase wobble between the mains at different parts of the country. The satisfactory results here are attributed to the less precise timing required in back-scatter work: back-scatter delays are 10–50 millisec compared, for example, with 0·3–1 millisec for the differences between various propagation modes over the path Slough–Inverness.

Table 1

TRANSMISSION SCHEDULES AND EQUIPMENT PARAMETERS

Dates of tests	11th–15th February inclusive	18th–22nd March inclusive
Rampisham transmitting schedule	0945–1000 G.M.T. 21·66 Mc/s 1600–1615 G.M.T. 21·64 Mc/s	0240–0245 G.M.T. 15·42 Mc/s. C.W. 0245–0300 G.M.T. 15·42 Mc/s. Pulse. (0245–0300 G.M.T. 647 kc/s, synchronized pulse transmissions)
Rampisham transmitter:		
Aerial input power	50 kW	50 kW
Pulse length	1 ms	1 ms
Pulse repetition frequency	25 pulses/sec	25 pulses/sec
Rampisham aerial*	Horizontal broadside array with reflector curtain (HR 4/5/0·8)	Horizontal broadside array with reflector curtain (HR 4/4/1·0)
Bearing of main lobe	200° true	184° true
Elevation of centre of lobe	7°	7·5°
Gain at centre of lobe with respect to $\lambda/2$ dipole in free space	20 dB	20 dB
Azimuthal beam width between 3 dB points	25°	25°
Slough schedule	Reception of Rampisham transmissions at above times. Transmission and reception for period following on nearby frequency.	
Slough transmitter:		
Aerial input power	100 kW	100 kW
Pulse length	120 μs	120 μs
Pulse repetition frequency	25 or 12½ pulses/sec	25 or 12½ pulses/sec
Slough aerial*	Horizontal rhombic (3-wire) $l = 212$ ft, $\phi = 70°$, $h = 85$ ft	Horizontal rhombic (3-wire) $l = 212$ ft, $\phi = 70°$, $h = 85$ ft
Bearing of main lobe	207° true	207° true
Elevation of centre of lobe	7·5°	11·5°
Gain at centre of lobe with respect to $\lambda/2$ dipole in free space	19 dB	16 dB
Azimuthal beam width between 3 dB points	14°	24°

* See Fig. 1 for vertical radiation pattern.

It was found during the tests that no ground-wave or forward-scatter pulse from Rampisham could be detected at Slough, so that no timing reference was available for measuring the delay of the back-scatter echoes. Meteor echoes were observed, but these occurred with varying time delays and were thus unsuitable. Various schemes for supplying a timing reference were tried unsuccessfully, and it was decided to arrange a further series of tests later when a medium-wave transmitter would be available to transmit a reference pulse.

No measurements of the range of the Rampisham back-scatter echoes were therefore possible, and attention was concentrated on obtaining measurements of echo amplitude and photographing the echo-pattern shapes. Range measurements were obtained on the back scatter from the Slough transmissions which followed the Rampisham schedule. The time-base was calibrated by marker pips at 200 and 1 000 km intervals for ease of range determination.

On account of the congestion in the broadcast bands, it was not possible to use normal wide-band pulse receivers for the present experiments. Satisfactory results were obtained by the use of a commercial receiver with modified i.f. output stage and a high-level detector.

The 1 millisec pulses could be received with a narrower bandwidth than the 120 microsec pulses, giving an enhanced signal/noise and signal/interference ratio.

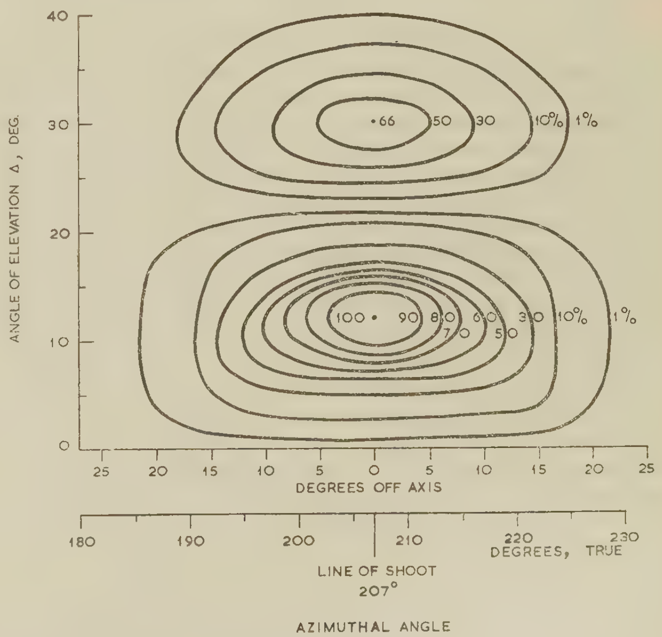


Fig. 2.—Radiation pattern of Slough rhombic at 15·42 Mc/s (March, 1957, test).

(2.2) Test on 18th–22nd March, 1957

The transmission schedule and equipment parameters relevant to this second test are also given in Table 1. A frequency of 15·42 Mc/s was used, so that, although the same receiving aerial was used at Slough as in the first test, its radiation pattern was altered. This radiation pattern is given in Fig. 2, which shows contours of constant power flux as a function of azimuth and elevation. This method of presentation is the one adopted by the C.C.I.R.[7]

The direction of shoot of the Rampisham aerial was 184° true and that of the Slough aerial 207°, a difference of 23°. The half-power beam widths were $\pm 12\cdot 5°$ and $\pm 12°$ respectively, so that for transmission from Rampisham and reception at Slough of long-distance echoes, the effective axis of the beam was at 196° true with a 6 dB loss relative to that for aligned beams.

During the period of the Rampisham pulse transmission a synchronized reference pulse was radiated on 647 kc/s, and this was received by ground-wave propagation at Slough. The phase of the mains-triggered 25 c/s receiver time-base was adjusted to place the Rampisham reference pulse near the beginning of the time-base and coincident with a 1 000 km range marker. When the receiver was tuned to 15·42 Mc/s the range of the back scatter could be read off from the range marks, taking as zero the 1 000 km marker previously identified. Occasional checks of the timing were made by tuning the receiver to 647 kc/s, but no relative drift between the time-base and reference pulse was found.

During this test a number of observers along (or close to) the great circle of the Rampisham aerial beam co-operated and sent reports on the signal strength of the Rampisham transmissions as received at their localities. The pulse transmissions were preceded by a five-minute c.w. transmission to enable 'SINPO' reports, of the kind customarily used in reporting broadcast reception (see Section 7.3) to be made. Reports were received from Cable and Wireless, Ltd., stations at Malta, Bathurst, Freetown, Lagos and Ascension Island, and from the Nigerian broadcasting station at Lagos. All the stations reported the signal strength of the c.w. transmission; Freetown, Malta, Lagos (Cable and Wireless, Ltd.) and Ascension Island measured the peak receiver input of the pulses in decibels relative to 1 μV, and the Nigerian broadcasting station at Lagos measured the relative amplitudes and delays of the received pulses.

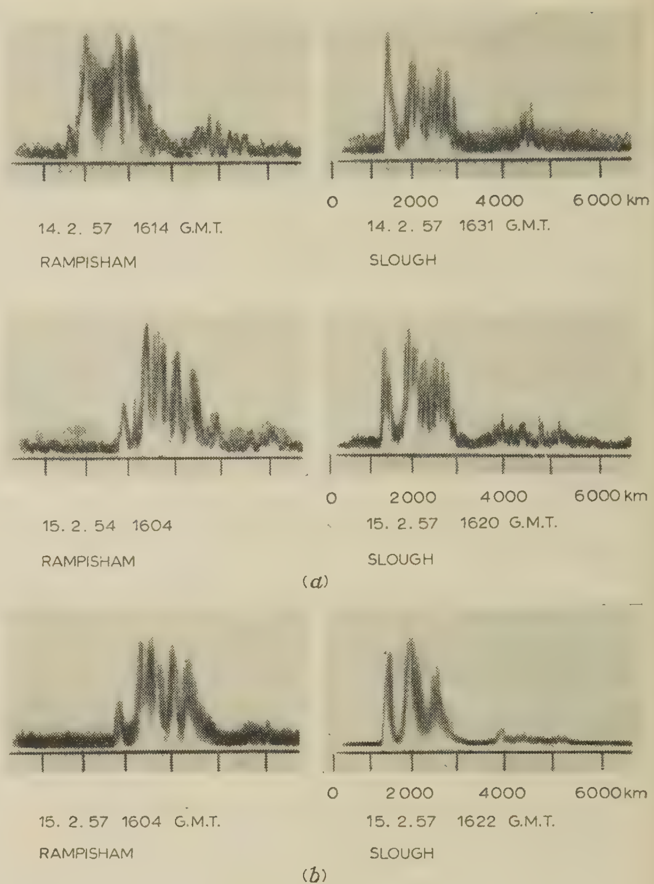


Fig. 3.—Echo patterns.

(*a*) Comparison of echo patterns obtained with Rampisham and Slough transmitting aerials (exposure 1 sec).
(*b*) Records of average envelope of echo patterns for Rampisham and Slough transmissions (exposure 30 sec).

(3) EXPERIMENTAL RESULTS

In the following discussion of the results, the differences between the echo patterns observed with the Slough and Rampisham transmitters, caused by the increased power and pulse width of the latter, are described first. The influences of aerial properties and propagation effects are then studied separately for the February and March tests. In the February test the propagation conditions were stable and the range of angles of elevation for which propagation was occurring were especially favourable for studying the influence of the aerial on the echo pattern. For this purpose, the patterns obtained with the Slough and Rampisham transmitters are compared with those calculated from ionospheric information by parabolic-layer theory and the radar equation.

In the March test, which took place at night, the propagation conditions were very variable and the results show the potentialities of back scatter for observing propagation conditions instantaneously. Theoretical calculations of pattern shapes are given to identify the 1-hop, 2-hop and 3-hop modes of propagation.

(3.1) Effect of Increased Power and Pulse-Width

The back-scatter echo patterns observed on a range/amplitude display during the transmissions from Rampisham showed markedly different characteristics from those normally observed with the Slough back-scatter sounder, this difference being attributable to the wider pulse used.

The Slough transmission, with a pulse width of 120 microsec, gave echo groups on the time-base in the form of 'mountain ranges', in which the individual peaks were narrow—of the order of the pulse width—and faded at random. Examples may be seen in Fig. 3(*a*) in the records marked 'Slough'. This random fading of a large number of peaks enabled the observer to average the envelope visually within a few seconds. Even the photographs mentioned, taken with a 1 sec exposure, give a good idea of the average pattern shape.

The Rampisham transmissions, for which the pulse width was 1 millisec, gave similar 'mountain range' scatter groups, but here the individual peaks were few in number and of the order of 1 millisec wide. Examples can be seen in Fig. 3(*a*) in the records marked 'Rampisham'. This coarser structure was much more difficult to average visually, because it was necessary to watch the pattern for a longer period to identify the envelope satisfactorily.

As a result of this coarse structure, the improvement in echo strength which resulted from the higher mean power of the Rampisham transmissions did not result in a corresponding improvement in visual detectability and ease of measurement of the echoes. The Rampisham transmissions were subject to interference on several of the observing days, but even when this was absent, some method of averaging the fading echo pattern was required to realize the full improvement in signal/noise ratio provided by the increased mean power.

The method adopted is illustrated in Fig. 3(*b*); a photographic exposure of 30 sec was employed, resulting in a thickened trace on the record, the densest part of the trace showing the most probable amplitude. Comparison of the Rampisham and

Slough records marked 15.2.57 in Figs. 3(*a*) and (*b*) shows the resulting improvement, particularly in the Rampisham records.

Measurements of the average amplitude of the peak of the first scatter group were made during all the test periods, the receiver gain being adjusted until the average echo strength (observed visually) gave a standard deflection on the cathode-ray-tube display. The measurements are summarized in Table 2.

Table 2

Date	Schedule	Receiver input voltage		Ratio of Rampisham input / Slough input
		Rampisham	Slough	
		μV	μV	
Feb. 11th, 1957	1600–1615	100	—	—
12th	0945–1000	—	72	—
	1600–1615	140	70	2·00
13th	0945–1000	105	73	1·44
	1600–1615	220	88	2·50
14th	0945–1000	90	28	3·22
	1600–1615	62	42	1·48
15th	0945–1000	56	—	—
	1600–1615	120	130	0·925
March 18th	0240–0245 (c.w.)	—	—	—
	0245–0300 (pulse)	174	41	4·24
19th	0240–0245 (c.w.)	340	—	—
	0245–0300 (pulse)	185	105	1·76
20th	0240–0245 (c.w.)	340	—	—
	0245–0300 (pulse)	68	79	0·86
21st	No observations			
22nd	0240–0245 (c.w.)	310	—	—
	0245–0300 (pulse)	180	102	1·77
			Mean, 2·02	

The signal-generator input voltage (in 75 ohms) to give the same deflection on the cathode-ray tube as the echoes is tabulated; this is a direct measure of the open-circuit voltage at the aerial terminals due to the received echoes.

The variations in the ratio (Rampisham input)/(Slough input) from day to day are seen to be considerable, probably partly due to the inaccuracies in the visual averaging method. The mean ratio, however, agrees with the expected factor of two deduced in Section 2. This exact agreement of the mean value is to some extent fortuitous in view of the small sample. Errors of 1–2 dB are also introduced by the inequality of aerial gains and beam-widths, and in the March test by beam misalignment.

Three measurements of the signal strength of the c.w. signals received from Rampisham in the March test are tabulated. The strength of these was greater than for pulses, as would be expected from the larger area of ground simultaneously illuminated, and also appeared to be more consistent.

(3.2) February Test—Effect of Transmitting-Aerial Properties

Inspection of Figs. 3(*a*) and (*b*) reveals one striking difference between the patterns obtained with the Rampisham and Slough transmitters. It will be noticed that both the patterns show a narrow peak followed by a trough at the beginning of the echo group, but the size of this peak relative to the later echoes is much greater for the Slough than for the Rampisham transmissions. Any such differences must be due either to differences in the transmitting aerial (the receiving aerial at Slough was the same in both conditions) or to the difference in the geographical location of the transmitter. The latter seems unlikely since the distance between the stations resolved along the line of shoot of the aerials is only 100 km, which is small compared with the ranges under consideration.

To investigate the effect of the aerials, the angle of elevation of the emitted and received energy must be found. This quantity can be determined from the oblique range and the height and ionization of the F2 layer, making use of the parabolic-layer theory[8,9] referred to in Section 7.1. Range measurement was not possible for the Rampisham transmissions, as explained in Section 2.1, but from the Slough record of Fig. 3(*b*) the oblique range of the first peak is 1 400 km, and of the first trough, 1 650 km. From ionospheric vertical soundings the ionization and height at the reflecting points have been estimated, and these yield angles of elevation for the peak and the trough of the echo pattern of 19° and 14° respectively.

Inspection of the vertical radiation patterns of the Slough and Rampisham aerials in Fig. 1(*a*) shows that an elevation of 19° is close to the angle of maximum radiation of the second lobe of both aerials, but the response is much less for the Rampisham than for the Slough aerial. An elevation of 14° is close to the first null for both aerials. The effect of the aerials on the scatter pattern is thus to reduce the amplitude of the leading edge, which would otherwise have the greatest amplitude, and to insert a 'notch' into the pattern. The smaller initial peak obtained with the Rampisham transmitting aerial is explained by the smaller response of its second lobe.

The effective vertical radiation patterns for transmission and reception at Slough, and for transmission from Rampisham and reception at Slough, have been calculated by multiplying the individual directivities and are shown in Fig. 1(*b*).

To identify the other features of the observed patterns, the pattern shapes for 15th February, 1600 G.M.T., have been calculated theoretically assuming that the ionization and height of the F2 layer varied along the path with the gradients predicted in propagation forecasts, but with the absolute magnitude of these quantities corrected by simultaneous measurements at Slough. (This procedure is described in more detail in Section 3.4.) Examination of the measurements of ionospheric characteristics made at ionospheric sounding stations along the path at 1600 G.M.T. on this day show close correspondence with this estimate. The lengths of hops have been found using the parabolic-layer relationships between angle of elevation and hop-length (Section 7.1) for the known layer height and ionization. Since the layer height and ionization vary along the path, the successive hops for a given angle are not necessarily the same length (see Fig. 4). For multi-hop propagation, therefore, the

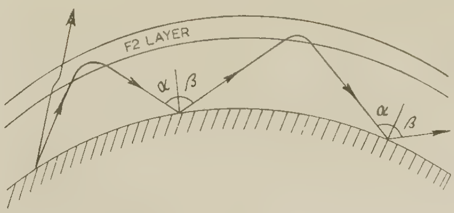


Fig. 4.—Effect of variations of layer height and ionization on lengths of successive hops.

technique is to match successive hops together so that the angle of incidence, α, and reflection, β, at the ground are equal, and the hop lengths correspond to those expected from the layer parameters at the successive reflection points in the layer. This

technique is due to Kift and has been described elsewhere.[5] The ground ranges calculated by this technique must then be corrected to obtain the equivalent free-space oblique path, p'.

The result of this calculation is shown in Fig. 5(*a*). The solid-line curves (mode plots) show the ground range plotted

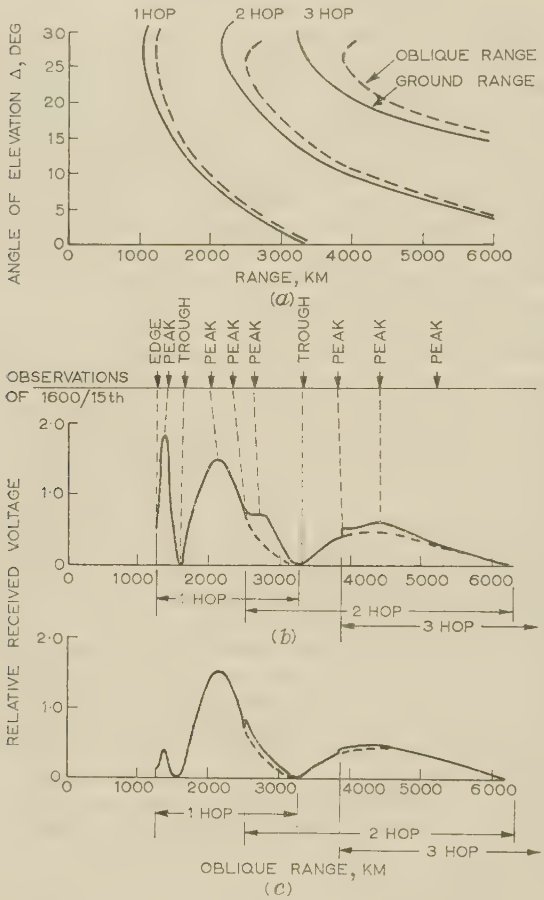


Fig. 5.—Curves for 1600 G.M.T., 15th February.

(*a*) Mode/angle plot.
(*b*) Calculated echo pattern for Slough transmitter. Characteristics of observed pattern, Fig. 3(*b*), included for comparison.
(*c*) Calculated echo pattern for Rampisham transmitter.

against angle of elevation for 1-, 2- and 3-hop rays. The dashed lines represent the corresponding oblique paths, and show the increased retardation suffered as the escape angle is approached.

This mode plot is applicable to the great circles through the axis of the Slough aerial beam, 207° true, and the effective axis for transmission at Rampisham and reception at Slough, 203°. In back-scatter reception, however, appreciable echo contributions are received from all directions within the main beam of the aerial, so that the mode plots for these directions should also be considered. Inspection of the ionospheric characteristics in the relevant area shows that the propagation pattern to distances of 6000 km from Slough should not have been significantly different over the $\pm 20°$ likely to be important.

With the relation between angle of elevation, Δ, ground range, D, and oblique path, p', provided by the mode plots, and the known aerial radiation patterns, the radar equation[9] can be applied to calculate the received back-scatter power as a function of oblique range (Section 7.2). The analysis yields the received power for each individual mode, whereas for certain ranges corresponding to certain time delays after the emission of the pulse, power is received simultaneously by two modes. Since the echoes by the two modes are in random phase, the powers can be added to yield the most probable value of power. The received voltage, which is the quantity observed on a range/amplitude display, is obtained by taking the square root of this total power. In Fig. 5(*b*) the solid line shows the calculated range/amplitude relation for 1600 G.M.T. on 15th February, the dashed lines indicating the voltage due to 1-hop and 2-hop modes alone.

In view of the quantities neglected in the analysis—skip-focusing, curved-layer focusing, 'time-focusing' and D-layer absorption—the agreement between the calculated pattern of Fig. 5(*b*) and the observed pattern of Fig. 3(*b*) (right-hand photograph) is remarkably good. The features of the observed pattern have been indicated above the calculated pattern for comparison, and are seen to coincide closely in range and to a lesser extent in amplitude. If skip focusing were taken into account, the leading edges of the three modes at 1000 km, 2500 km and 3800 km would have a considerably enhanced amplitude corresponding more closely to the narrow peaks observed. The inclusion of curved-layer focusing might explain the enhancement of the 2-hop echoes relative to the calculated values.

In Fig. 5(*c*) the pattern shape calculated for transmission from the Rampisham aerial is given for comparison with the left-hand photograph of Fig. 3(*b*). The initial peak is much smaller than in Fig. 5(*b*) as a result of the smaller second lobe of this aerial, as discussed above. The remaining features of the pattern are similar to those calculated, except for the first part of the 2-hop echo, which is barely distinguishable on the calculated pattern, but clearly visible on the photograph. This difference may be exaggerated by the power-addition technique which accentuates errors made in calculating the weaker of two echoes.

In general, however, the calculations indicate that the back-scatter pattern can be explained, even when it is complex and composed of several modes, and the results suggest that the observed patterns could be used to determine coverage and aerial performance.

(3.3) March Test—Indication of Propagation Conditions

This test was conceived as a check on the accuracy of back-scatter observations for determining the illuminated areas of the earth. Skip distance measurement by back scatter has been studied previously and its accuracy and limitations are known, but little work has been done on the accuracy of interpretation of multi-hop echoes.

Back-scatter observations in the past have often shown 1- and 2-hop echoes with a gap between them, indicating that the illumination is not constant along the line of shoot but is concentrated by vertical aerial directivity and skip focusing on to certain areas of ground. The occurrence of a gap between the back-scatter echo groups on the time-base means that the illumination between the intensely illuminated zones is insufficient to give detectable echoes, although it may be sufficient for communication. However, on some occasions when the echo strength has been very high, there has still been a gap between the echo groups, this gap appearing when the ratio of transmitted frequency to critical frequency exceeds about 2·0, and widening with further increase in frequency. This effect would be expected if specular reflection takes place at the ground and if there is no marked scattering from any part of the ionosphere. Fig. 6 shows the ray configuration; the skip ray, after reflection at the ground, returns after a further hop to a point beyond the limiting range for one hop.

In the March test it was hoped to choose a frequency for which such a gap would occur, the gap embracing some of a

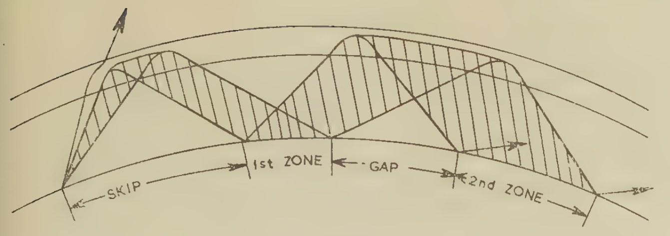


Fig. 6.—Formation of gap between the first and second illuminated zones.

chain of listening stations which would report on the received field strength. If very low field strengths were reported by stations at the location of the gap, as determined by back-scatter observations, and high field strengths by stations on either side, then confirmation of the accuracy of the techniques would be provided.

Unfortunately the critical frequencies were too high for the purpose during the period, and although a small gap appeared it did not embrace any of the listening stations. However, although this particular test was unsuccessful, other observations provided by the listening stations confirmed the accuracy of the

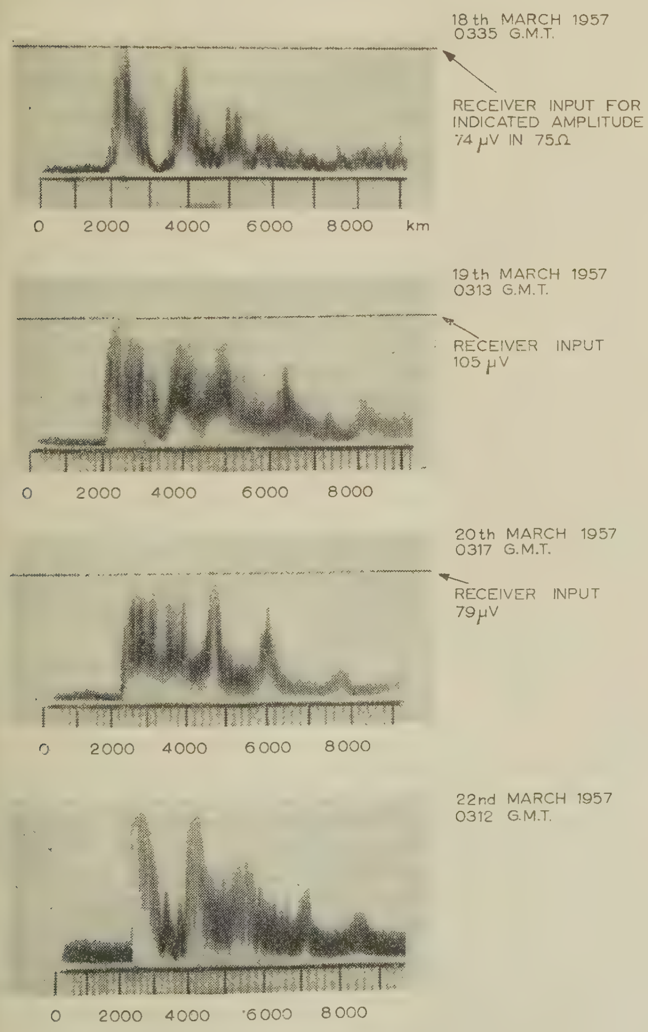


Fig. 7.—Back-scatter echo patterns, Slough transmitter, 18th–22nd March.

back-scatter indications, and the experiments demonstrated the usefulness of the technique for indicating day-to-day changes in propagation conditions.

Photographs of the range/time display for the Slough transmissions are given in Fig. 7 for each day of the test except 21st March, when no observations were made at Slough. The photograph for 18th March was taken with a 1 sec exposure, the rest with a 30 sec exposure. For the 18th, 19th and 20th, reference lines are drawn on the display showing the deflections corresponding to the indicated open-circuit input voltages; these were found by signal-generator calibrations.

The results obtained during the Rampisham transmissions were generally similar to these Slough observations, but the patterns were obscured by 'echoes round the time-base'. The reason for this can be seen from Fig. 7, in which echoes are recorded at ranges up to 9000 km; this corresponds to a delay time of 60 millisec, or $1\frac{1}{2}$ periods at 25 c/s, so that echoes from one pulse were still being received after the next pulse had been radiated. For the Slough transmission $12\frac{1}{2}$ c/s was employed, enabling the full range of echoes to be recorded without overlap.

The patterns are seen to consist of a series of peaks which correspond to the ground-reflection points for multi-hop propagation. The patterns are of simpler structure than for the February test, since the skip ray was at an angle of elevation smaller than that of the first null of the aerial radiation pattern (22° from Fig. 2). There was thus no 'notch' in the 1-hop echo group, and this and the multiples show the more normal sharp leading edge followed by a gradual trailing part. There was a small variation from night to night in the range of the leading edge of the first group, ranges from 2000 km to 2400 km being observed, but the variations in the multiple hops were much larger.

(3.4) March Test—Identification of Propagation Modes

In order to identify the modes corresponding to the various features of the observed echo-patterns, mode plots, showing the angles of elevation for propagation to any distance along the great circle of the Slough aerial beam, have been calculated by the theoretical methods of Section 7.1, as for the February test.

For a post-mortem examination of propagation conditions, such mode plots can be based on measured ionospheric characteristics when, in due course, these have been received from ionospheric sounding stations along the path. However, these data are not available at the time of the back-scatter observations, and use must be made of world maps of predicted F2-layer critical frequency and 4000 km m.u.f. (The height of maximum ionization of the F2 layer, h_m, a required parameter, may be derived from the maps by taking the ratio (m.u.f. 4000 km)/(F2 critical frequency) for points along the path and determining the corresponding value of h_m by means of a conversion chart.[4])

Accordingly, the mode plots of Fig. 8 have been calculated by taking from the 'Predictions of Radio Wave Propagation Conditions'[10] for March, 1957, the predicted variation of critical frequency and layer height as a function of distance from Slough along the path. For each night of the test, the critical frequency at every point along the path has been multiplied by the factor necessary to make the critical frequency at Slough equal to the measured value. This procedure preserves the predicted gradient but corrects the absolute values, at least at the Slough end. It was found unnecessary to correct the predicted layer heights at the Slough end since the measured values did not vary much.

Subsequent investigation of measured ionospheric characteristics at Dakar and Casablanca for these days and times showed differences between the corrected predictions and the measured

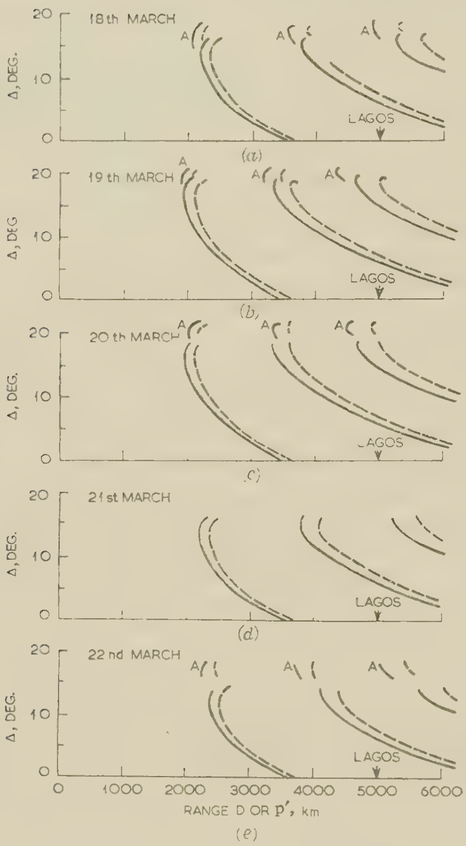


Fig. 8.—Mode plots for March experiments.

——— Ground range, D.
— — — Oblique equivalent path, p'.
Complete curves are from ionospheric predictions corrected by vertical-incidence measurements at Slough only.
Curves A are from vertical-incidence measurements along whole path.

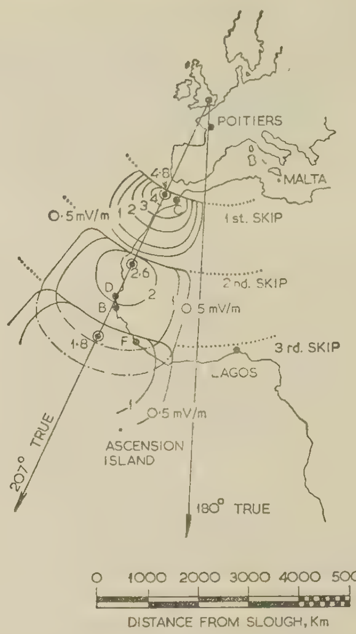


Fig. 9.—Field strength set up at ground by ionospherically-reflected radiation from rhombic aerial at Slough.

From 0300 G.M.T. March predictions.
—·—·— Contours for the 2-hop field at locations for which the 3-hop field is stronger.
F Freetown. B Bathurst. D Dakar. C Casablanca.

values, indicating that the gradient varied from night to night. Sections of the mode plots derived from these measured characteristics are given in Fig. 8 at the points marked A showing the minimum distances for the various modes. The measurements for the 19th and 20th March agreed fairly closely with the predictions, at least for the first and second hops.

It is of interest to compute from these mode plots the manner in which the energy radiated by the Slough transmitter was distributed over the service area. Fig. 9, which gives the results of such a calculation, shows the skip-distance contour, and also the contours of 'second skip' and 'third skip' (more strictly, the minimum distances for 2-hop and 3-hop propagation). These contours are derived by producing mode plots (approximately representing the conditions of March 19th) for a number of directions of propagation so as to determine how the skip distance changed over the aerial beam. Fig. 9 shows that this variation is small.

To determine the field incident on the ground in the area due to radiation from the Slough transmitter, the angle of elevation for propagation to a particular distance has been found from the mode plot of Fig. 8(*b*), and the corresponding aerial gain found from the radiation pattern of Fig. 2. Taking a transmitter power of 100 kW, an aerial gain at the lobe maximum of 16 dB relative to a free-space dipole, and assuming an inverse-distance field, the contours of incident field in millivolts per metre have been calculated and are plotted in Fig. 9. The Figure shows a drop in field strength to 0·5 mV/m at the far edge of the first illuminated zone at 3 200 km, rising rapidly to 2·6 mV/m at 3 500 km where the ground is illuminated by 2-hop rays at the angle of elevation of the lobe maximum. The 2- and 3-hop zones overlap to a considerable extent, resulting in a much smaller enhancement.

To provide a comparison for the observed back-scatter echo patterns of Fig. 7, theoretical echo patterns have been calculated from the mode plots of Fig. 8 in the same manner as for the February test. The calculated patterns are shown in Fig. 10, together with arrows indicating the positions of the leading edges of the observed echo groups.

The correlation of the calculated and observed leading-edge positions is quite good for the first hop, but poorer for the later hops, the observed ranges being, in general, shorter than those predicted. It is significant that the best agreement is for the 19th and 20th March, the days for which the measured ionospheric characteristics along the path were closest to the predicted values. Even on these occasions, however, the third echo group was short of the predicted range.

To find out whether the observed echo patterns agreed with those expected from the actual ionospheric conditions along the path at the times of the observations, ionospheric characteristics measured at Poitiers, Casablanca, Dakar and Ibadan on the nights of the tests were examined. From these the true gradients of layer height and critical frequency were estimated, and mode plots were calculated. The resulting corrections to the leading edges of the echo groups are shown at A in Figs. 8 and 10. The agreement between observed and calculated leading-edge ranges is much closer than for the corrected predictions.

Study of the ionospheric measurements at Casablanca and Dakar also reveals the reason for the night-to-night variation in the echo pattern. The ratio (f_0F2 at Dakar)/(f_0F2 at Casablanca) for 0300 G.M.T. in March, 1957, varied from 0·7 to 1·5 of the mean value, showing that the horizontal ionospheric gradient was very variable.

The conclusion drawn from these comparisons is that the

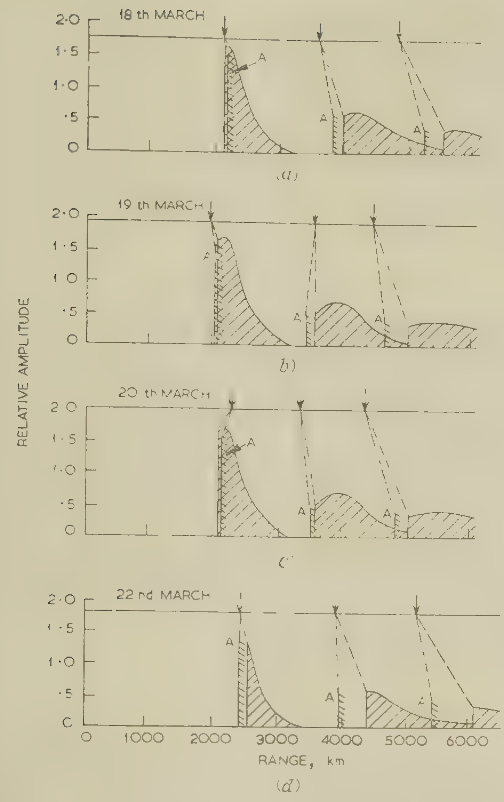


Fig. 10.—Comparison of observed and calculated echo patterns.

Full shaded areas.—Calculated from ionospheric predictions corrected by vertical soundings at Slough.
A.—Calculated from vertical soundings along the complete path.
Arrows.—Leading edges of observed echo groups (Fig. 7).

back-scatter measurement indicated the propagation modes over the 6000 km path with much better accuracy than that obtainable by extrapolation from vertical-incidence measurements at one terminal. The vertical-incidence measurements along the path showed the large variation encountered at night and explained the changing character from night to night of the multi-hop scatter patterns of Fig. 7. Variations in the daytime were much smaller, leading to the more consistent results observed in the February test.

(3.5) March Test—Listeners' Reports

Inspection of the photographs in Fig. 7 shows that on none of the nights did a gap exist between first and second echo groups which was wide enough to embrace any of the listening stations. The reports from these stations (Table 3) showed consistently strong signals, as expected from the scatter results and predictions. An exception was Malta, 2100 km distant from Slough, which gave much lower signal strengths on 20th and 22nd than on 19th and 21st. The skip distance deduced from the range of the leading edge of the back scatter on the 20th and 22nd are slightly greater than 2100 km, placing Malta in the skip zone, while those for 19th and 21st are slightly less than 2100 km, placing Malta in the illuminated zone. The deductions from back scatter are thus in agreement with the signal-strength observations. Not too much weight should be attached to this result, however, because the bearing of Malta is 60° off the axis of the aerial, giving a distance between the reflection points at the ionosphere of 1000 km.

A further interesting observation was made by the Nigerian broadcasting station observers at Lagos. They reported that two major pulses were received, separated in time by something less than a millisecond, with the second pulse of slightly less amplitude than the first. Fig. 7 shows that the angles of elevation expected at Lagos (5000 km distant) were about 6° for the 2-hop mode and 13–16° for the 3-hop mode, the latter angle

Table 3

Report on B.B.C. Transmissions 18th–22nd March, 1957 (15·42 Mc/s)

Receiving station	Aerial	March 18th		March 19th		March 20th		March 21st		March 22nd	
		SINPO	Peak pulse*	SINPO	Peak pulse*	SINPO	Peak pulse*	SINPO	Peak pulse*	SINPO	Peak pulse*
Ascension Island ..	London receiving HAD (18 Mc/s)	44434 Fading period 3–7 sec	dB +58	44434 Fading period 3–7 sec	dB +65	44434 Fading period 3–7 sec	dB +60	44434 Fading period 3–7 sec	dB +67	44434 Fading period 3–7 sec	dB +67
Malta	—	—	—	45444	+72	35423 Rapid fading	+45	44433 Very rapid fading at commencement	+70	34523 Rapid fading	+26
Freetown	Aperiodic L	—	+46	—	+44	—	+40	—	+45	—	+52
Lagos (C. and W.) ..	London receiving rhombic	— Fading period 15–30 sec	+66	—	+64	— Fading period 15–30 sec	+70	— Fading period 3–5 sec	+58	— Fading period 5–10 sec	+56
Lagos (N.B.S.)	—	—	—	SIM 545	—	SIM 545	—	—	—	SIM 545	—
Bathurst	—	No measurements made as interference from French- and American-speaking stations made pulses on oscilloscope very variable. Signal strength QSA5 and would be good but for this interference.									

* In decibels relative to 1 μV input to receiver.

being more critically dependent on the layer height and ionization.

The gain of the Rampisham aerial for the higher-angle mode is seen from Ref. 7 to be considerably less than for the lower-angle mode, so that the near equality of the amplitudes of the two components observed at Lagos would not have been expected on these grounds alone. However, the receiving aerial at Lagos was inside a building and of low height, and is thus likely to have had a poor pick-up at the lower angle. This may have resulted in the relative enhancement of the higher-angle mode, making the two amplitudes comparable, but it is not possible to make a quantitative estimate of this effect.

The difference in time of arrival for the two modes is estimated as about 0·7 millisec, which is in accord with the observations.

The pulse-reception reports from Lagos provide one important confirmation of the propagation deductions made from the back-scatter observations, which to some extent offsets the failure of the original conception of an all-or-nothing test of the usefulness of back scatter for locating illuminated zones. The observers report that two major pulse components, as described above, were seen on each night of observation, these nights being the 19th, 20th and 22nd March. On the 22nd March the calculated predictions of Fig. 8(*e*), obtained by extrapolating from local measurements at Slough, indicate propagation to Lagos by two hops only, the 3-hop rays returning to the earth beyond Lagos. The back-scatter observations, shown by the arrows in Fig. 10(*d*), indicate that 3-hop propagation should have been possible for distances equal to or greater than 4 850 km (oblique range $\geqslant$ 5 200 km) and thus suggest that two modes should have been received at Lagos on this night. The mode plot derived from actual ionospheric measurements [A in Fig. 8(*e*)] also indicates 2- and 3-hop propagation to Lagos.

The Lagos observation thus demonstrates that propagation conditions to a point 5 000 km distant were correctly evaluated by scatter sounding, at a time when extrapolation from local measurements by means of the predicted ionization gradient gave an erroneous result.

(4) CONCLUSIONS

The two series of tests described have demonstrated the potentialities of scatter sounding, both as an engineering tool for finding the most effective frequency and aerial system to obtain broadcast coverage of a specific geographical area, and as an instantaneous indicator of propagation conditions.

The February test showed that under stable ionospheric conditions the features of the observed echo-patterns, although of a complex nature in this case, were in accord with those calculated from aerial characteristics and measured ionospheric conditions. In this test the difference in the coverage produced by the two aerial systems could be clearly seen. The results suggest the use of back scatter for determining rapidly the optimum vertical radiation pattern required for covering a particular area.

In the March test, the reports received from listeners along the propagation path were in accord with the deductions made from the back-scatter echo-patterns, showing that the techniques could provide a useful instantaneous indication of the propagation modes effective at all distances along the path.

These instantaneous measurements of propagation conditions to points 5 000 km distant may be compared with the only other method of propagation evaluation possible from one terminal of a path. This method, an extrapolation procedure, is to assume that the ionospheric height and critical frequency vary along the path with the gradients derived from ionospheric prediction maps, but with the absolute values corrected so that the height and critical frequency at the observing terminal correspond with simultaneous measurements using a vertical-incidence ionospheric sounder.

The reports from Malta in these tests suggest that the skip distance deduced from back scatter was accurate, while the observations of pulse patterns at Lagos confirmed the propagation modes deduced from back scatter, although these were contrary to the modes expected from the extrapolation procedure. Further confirmation of the accuracy of modes deduced from back scatter was provided by their agreement with those calculated from simultaneous vertical ionospheric soundings along the path.

(5) ACKNOWLEDGMENTS

The tests were carried out at the request of the British Broadcasting Corporation, and thanks are due to Mr. T. W. Bennington and to the engineers of the External Broadcasting Department for their close co-operation throughout.

The listeners' reports in the second test were provided by the engineers of a number of stations of Cable and Wireless, Ltd., through the agency of Mr. R. J. Hitchcock of that Company, and also by the Engineer-in-Charge of the Nigerian Broadcasting Station at Lagos.

The work was carried out as a part of the programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(6) REFERENCES

(1) VILLARD, O. G., and PETERSON, A. M.: 'Scatter-Sounding: A Technique for the Study of the Ionosphere at a Distance', *Transactions of the Institute of Radio Engineers*, 1952, **PGAP–3**, p. 186.
(2) SILBERSTEIN, R.: 'Sweep-Frequency Backscatter—Some Observations and Deductions', *ibid.*, 1954, **AP–2**, p. 56.
(3) SHEARMAN, E. D. R.: 'A Study of Ionospheric Propagation by Means of Ground Back-Scatter', *Proceedings I.E.E.*, Paper No. 1914 R, October, 1955 (**103 B**, p. 203).
(4) WILKINS, A. F., and SHEARMAN, E. D. R.: 'Back-Scatter Sounding: an Aid to Radio Propagation Studies', *Journal of the British Institution of Radio Engineers*, 1957, **17**, p. 601.
(5) KIFT, F.: 'The Propagation of High-Frequency Waves to Long Distances', *Proceedings I.E.E.*, Paper No. 3156 E, March, 1960 (**107 B**, p. 127).
(6) SHEARMAN, E. D. R., and MARTIN, L. T. J.: 'Back-Scatter Ionospheric Sounder', *Wireless Engineer*, 1956, **33**, p. 190.
(7) C.C.I.R. 'Antenna Diagrams', International Telecommunication Union, Geneva, 1953.
(8) APPLETON, E. V., and BEYNON, W. J. G.: 'The Application of Ionospheric Data to Radio-Communication Problems: Part I', *Proceedings of the Physical Society*, 1940, **52**, p. 518.
(9) SHEARMAN, E. D. R.: 'The Technique of Ionospheric Investigation using Ground Back-Scatter', *Proceedings I.E.E.*, Paper No. 1915 R, October, 1955 (**103 B**, p. 210).
(10) 'Predictions of Radio Wave Propagation Conditions. Bulletin A.' (Radio Research Station, Slough, England.)

(7) APPENDICES

(7.1) Calculation of Angle of Elevation as a Function of Distance for a given Ionospheric Configuration

Appleton and Beynon[8] (see also Shearman[9]) give the relation between angle of elevation, Δ, and ground-range, D, for a ray reflected from a parabolic layer as

$$D = D_1 + D_2$$

$$= \frac{R}{R + h_0} \sin i_0 2xy_m \tanh^{-1}\left(\frac{x \cos i_0}{1 - \frac{x^2 y_m}{R + h_0} \sin^2 i_0}\right)$$

$$+ 2R\left(\frac{\pi}{2} - \Delta - i_0\right) \quad . \quad (1)$$

where

$$i_0 = \sin^{-1}\left(\frac{R}{R + h_0} \cos \Delta\right) \quad . \quad . \quad . \quad . \quad (2)$$

From this relationship, curves of D plotted against x for various values of h_m have been prepared.

If h_m and x for the path under consideration are plotted against D, the length of the first hop for a chosen angle may be found by a superposition technique. The length of the second hop is then found by assuming specular reflection at the ground so that the same value of Δ is maintained, using the superposition technique again to find the hop length.

This procedure, due to Kift,[5] has been adopted to plot the D/Δ curves of Figs. 5(*a*) and 8.

In back-scatter sounding the observed quantity is not ground range but oblique equivalent path, p'. This is related[9] to D by

$$p' = p_1' + p_2 = \frac{D_1}{\sin i_0} + \frac{2R \sin (\pi/2 - \Delta - i_0)}{\sin i_0} \quad . \quad (3)$$

From eqn. (3) the p'/Δ curves shown dotted in Figs. 5(*a*) and 8 have been plotted.

(7.2) Calculation of Theoretical Back-Scatter Range/Amplitude Relationship from the Radar Equation

The radar equation gives the echo power received from a target at range r in free space as

$$P_R = \frac{P_T G_T}{4\pi r^2} \frac{A}{4\pi r^2} \frac{G_R \lambda^2}{4\pi} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

The echoing area is defined as 'the aperture which, if placed normal to the wavefront incident upon the target, would intercept an amount of energy that, radiated isotropically, would give rise to the observed incident echo flux at the receiver'.

In applying this equation to the ground back-scatter problem we take r to be the oblique path from the transmitter to the distant ground irregularities by way of the ionosphere (see Fig. 11). In the following calculations the equivalent free-space path, p', given by the free-space velocity of light times the time of travel, is taken instead of the true path; this is the appropriate value for ionospherically refracted rays if focusing can be neglected.

For back scatter from distributed targets such as the ground irregularities of interest here, the echoing area will be dependent on the ground area simultaneously illuminated. This ground area is that subtended in azimuth by the aerial beam, and in range by two rays differing in length by $\frac{1}{2}c\delta t$.

From Fig. 11, the ground area illuminated is

$$R \sin \frac{D}{R} \theta_B \tfrac{1}{2}c\delta t \sec \Delta \quad . \quad . \quad . \quad . \quad . \quad (5)$$

This is the total area of ground illuminated simultaneously; to find its echoing area the scattering properties of the ground must be known. In the absence of experimental data at these wavelengths, it is assumed here that a fraction $1/K$ of the energy incident upon the ground is scattered isotropically, the rest being specularly reflected. This assumption is equivalent to putting the echoing area A in eqn. (4) equal to $1/K$ times the projected ground area. From Fig. 11 this projected area will be $\sin \Delta$ times the ground area. Thus, from eqn. (5),

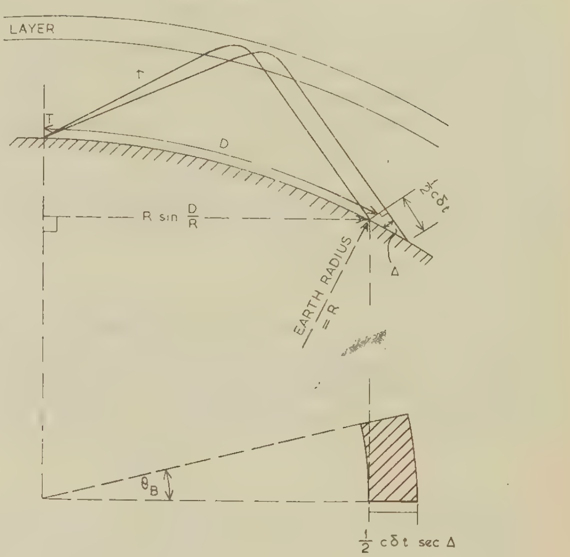


Fig. 11.—Ground area simultaneously illuminated by a pulse transmission.

$$A = \frac{1}{K} R \sin \frac{D}{R} \theta_B \tfrac{1}{2}c\delta t \tan \Delta$$

Substituting for A in eqn. (4),

$$P_R = \frac{P_T G_T G_R \lambda^2}{64\pi^3 r^4} \frac{1}{K} R \sin \frac{D}{R} \theta_B \tfrac{1}{2}c\delta t \tan \Delta$$

In this study we are interested only in the relative received power, so that constant terms may be disregarded. Hence

$$P_R \propto \frac{G_T G_R}{r^4} \sin \frac{D}{R} \tan \Delta \quad . \quad . \quad . \quad . \quad . \quad (6)$$

In this equation the power gains of the transmitting and receiving aerials, G_T and G_R, are given as a function of Δ in Figs. 1(*a*) and 2. (The relative gain with respect to that at the centre of the lobe is plotted in the Figures; this is all that is required for the present purpose.) The other quantities required are r, D and Δ, and the relationship between these can be found for specified ionospheric conditions and transmitted frequencies from the Appleton and Beynon parabolic layer theory, as shown in Section 7.1.

The procedure for calculating the relationship between echo amplitude and the range from these equations is as follows:

(*a*) From the parabolic-layer theory and the ionospheric configuration along the path, the angle of elevation, Δ, of 1-hop, 2-hop and 3-hop rays as a function of distance, D, is found and the corresponding oblique equivalent path, p', is deduced.

(*b*) From (*a*) and eqn. (6) the relative received power, P_R, as a function of distance D for each mode is found.

(*c*) Using the parabolic-layer-theory eqns. (1), (2) and (3), the received-power/distance relationship for each mode is converted to a received-power/range relationship, taking range r as being equal to the equivalent path p'.

(*d*) The powers received from a given range for the three modes of propagation are added, and the square root of the sum is taken. Since the energy received by the different modes will be in random phase, this method will yield the most probable value of the voltage at the receiver terminals. The range/voltage curve so calculated is the required pattern shape as seen on a range/amplitude cathode-ray-tube display.

This analysis neglects the focusing of energy introduced by the thickness of the layer, an effect which considerably enhances the energy incident near the skip distance. Focusing due to the curvature of the layer also gives an enhancement of the energy incident on the ground, this being more marked as the tangential-ray condition is approached. For daytime transmission ionospheric absorption should also be considered in the calculation. These effects will distort the range/amplitude patterns calculated by the simplified assumptions above, as discussed in Section 3.2.

(7.3) SINPO Code Used in Reporting Broadcast Reception

SINPO is a 5-figure code in which figures taken from the tables below are grouped in the order S, I, N, P, O to give an overall rating of the quality of the received signal.

S = Carrier strength	Excellent	5
	Good	4
	Fair	3
	Poor	2
	Barely audible	1

I = Interference	Nil	5
	Slight	4
	Moderate	3
	Severe	2
	Extreme	1

N = Atmospheric noise	Nil (−40 dB)	5
	Slight (−30 dB)	4
	Moderate (−20 dB)	3
	Severe (÷10 dB)	2
	Extreme (0 dB)	1

P = Propagation disturbance	Nil	5
	Slight	4
	Moderate	3
	Severe	2
	Extreme	1

O = Overall readability	Excellent	5
	Good	4
	Fair	3
	Poor	2
	Unusable	1

DISCUSSION BEFORE THE ELECTRONICS AND COMMUNICATIONS SECTION, 6TH MARCH, 1961

Mr. J. K. S. Jowett: The potential value of back-scatter sounding as an operational tool for commercial services first began to be discussed some 10 years ago, but it appears that the early promise shown by those techniques have not been realized, so far, in everyday practice. The paper presents a very good opportunity for asking ourselves why this is so.

The investigation which it describes relates only to its possible use for broadcasting services in the h.f. band and this is clearly the most obvious application. While the discussion will doubtless centre on this application, I hope that the possible application of back-scatter techniques to other radio services will also be considered.

The author's conclusions are not, perhaps, as clear and definite as one might wish; for what types of routes and conditions does he think these conclusions can safely be drawn—for one- or two-hop paths only, or for longer paths? Would they apply equally for east–west routes, which are more difficult, as well as for north–south routes such as those chosen for the test?

It is generally recognized that, on fixed services, back-scatter as an aid in frequency selection is unlikely to be of any great additional help. This is because at present such services have to be supervised by technical operators who, amongst their other duties, are generally able to decide the best frequency to use at any particular time.

In mobile services, and in particular the aeronautical services, there may well be more room for the application of back-scatter for help in frequency selection.

From the point of view of a fixed service, the most stimulating suggestion in the paper is that back-scatter sounding can help to indicate the optimum aerial system to be used for a particular service. This might mean that back-scatter sounding could become more important as a planning tool than as an operational aid. For example, we in the Post Office are making long-term measurements of the arrival angles of high-frequency signals observed over a number of routes in order to take fullest advantage of the prevailing modes of propagation. These measurements involve the provision of pulsed transmissions from the distant terminal. We have been fortunate in obtaining the excellent co-operation of several administrations in this, but an extension to many routes would be most difficult to arrange. Is it possible that back-scatter sounding could be a cheap, do-it-yourself method for finding out the answers without the need for co-operation from a distance? Results which we have already obtained from a series of back-scatter tests over an Atlantic path illustrate the difficulties which arise in interpretation. How does the author suggest these difficulties might be met in an operational system?

There is also the sensitivity problem of how to make back-scatter applicable to longer-distance systems. In our tests we are now planning to use a frequency-modulation technique for back-scatter; by this means we should utilize more fully the 30 kW of c.w. power available at the transmitter and gain considerably in sensitivity and range. Has the author made any use of such a system and does he agree that some such means of increasing sensitivity and range would be necessary for operational use?

Mr. N. G. V. Anslow: Because the radiated power from an aircraft is small, with a theoretical maximum of 100 W, our operation is essentially marginal as regards communication and so can stand very little degradation of the transmission path.

Each aircraft as it departs is given a primary and a secondary frequency on which watch must be kept and contact made with the ground station. The primary frequency is based upon prediction; the secondary frequency, which will differ by as much as 3 or 4 Mc/s, will, generally speaking, be lower. These two frequencies will be used over a wide range of 350–1 500 n.m. from the ground station unless circumstances necessitate a change. On the North Atlantic the frequency bands are further subdivided into families of frequencies within the band so that, for example, eastbound and westbound aircraft can be segregated. This has become necessary because of the high density of traffic in this area.

In B.O.A.C. trials last year we found that, by using a back-scatter-measured frequency as opposed to a predicted frequency, we were frequently able to achieve a very significant improvement in communications since the optimum frequency for the known distance of the aircraft could be chosen. We found cases where aircraft in some areas could operate on frequencies much higher than the maximum indicated by propagation predictions. This is doubly advantageous, for, apart from getting aircraft on to the optimum working frequency and hence achieving a highly satisfactory communications link, it reduces

the numbers of aircraft operating on the originally predicted frequencies for the area: hence the congestion is reduced with a consequent improvement in traffic handling.

Mr. R. J. Hitchcock: On looking at the echo patterns in the paper I am immediately reminded that to provide a 'wanted' signal over the few acres of land on which some distant antenna stands, great areas of the earth are covered with 'unwanted' signals of equal or greater intensity; no wonder the h.f. spectrum is plagued with interference. While the broadcast nature of h.f. is inevitable, it is obviously desirable that the area of maximum field should, as far as is practicable, include the receiving location. At present we have no idea whether the maximum field lies short of, on, or beyond the receiver. In back scatter, however, we have a technique which shows not only this area but also its movements with changes in frequency and/or antenna. The requirement now is for an inexpensive equipment channelled to the normal transmissions—an operational meter—so that with judicious use of available frequencies and antennae, an area of maximum field may be steered to the required location. The output power could then be reduced and the interference potential decreased.

I was particularly interested in the gap between the first and second hops, but what further work is planned to confirm whether the signals are sufficiently strong for commercial reception where echoes may not be detectable? This is important as the potentialities of back scatter are greater for the shorter routes, many of which must have receiving locations in or near this particular gap. Also it would have been interesting to see summer experiments when E and Es might have been significant.

Finally, it may be disquieting for operating organizations who hope later to use propagation modes in their predictions to know that such modes may be inaccurate when computed by the accepted method, namely from vertical-incidence data.

Mr. P. A. C. Morris: In working 24-hour point-to-point circuits it is not, in general, difficult to maintain contact for 20 hours out of the 24. It is the odd 4 hours that merit our attention and for which we are now turning to ionospheric soundings, whether by back scatter or by oblique-incidence swept-frequency techniques. This period coincides with the pre-dawn ionization dip, at which time senior engineering staff are not on duty, so that in applying these techniques we have a staffing problem as well as an engineering one.

Does the author consider the system antenna front-to-back ratio to be adequate to discriminate against backward echoes, and in particular against auroral echoes?

Cable and Wireless Ltd. at first used a separate pulse transmission in tests between Singapore and Hongkong, but since this occupied valuable equipment, efforts were made to pulse on the spare sideband of the normal multiplex telegraph circuit. The technique is to apply a 4 kc/s audio pulse to the i.s.b. drive equipment and in this way to gain the benefit of the i.s.b. audio filters. This has been very successful in that indication of the fade-out time at the distant receiving station has been obtained with an accuracy of a few minutes. Although the peak pulse power is only about 1 kW, we were gratified to obtain echo ranges up to 5 000 km.

We have also experimented with the swept-frequency continuous-wave ranging system described by Hymans and Lait.* A saw-tooth waveform is applied to the reactance valve of a diplex keyer which modulates a standard transmitter. The time delay between the ground wave and the back-scatter echo appears as a beat tone which is displayed on a spectrum analyser. This method gives an improved signal/noise ratio but is rather impractical at present because of the amount of equipment required.

Mr. J. H. H. Merriman: In h.f. point-to-point radio systems it is obviously profitable to remove all unnecessary man-power and skill from transmitting and receiving stations by installing plant of high intrinsic reliability which can be remotely controlled. The proposed techniques, however, create a need for a high degree of skill in interpretation, for increased complexity in plant at the transmitting stations and for some further increase in plant at receiving stations.

Does the author see any way of simplifying the interpretation and the apparatus associated with back scatter; does he see a way of mounting a back-scatter operation in a transmitting station, with the associated reception perhaps going on at the same place, the data being transmitted back in some way to the control-point for the service (which point, in general, will not be a transmitting or receiving station but well removed from both)?

Secondly, what real freedom exists to make use of the data which back-scatter may provide? The International Frequency List for the frequency range 4–27·5 Mc/s at the present time contains some 170 000 registrations in that band. It is clear that this very large number, of itself, prevents the potential of back scatter from being developed. Could the author comment on the actions he might expect operating authorities to take on receipt of this additional information? Are the actions likely to yield significant improvement in operational performance and to offset the considerable cost and complexity involved?

Lastly, what is the interference potential of back scatter, both in terms of interference to other services (can, in fact, sufficient resolution be provided with pulse shapes such that the radiated energy is contained within a 6 kc/s channel) and in terms of the extent to which live channel time is to be devoted to non-traffic transmissions?

Dr. G. J. Phillips: When investigating parameters that affect the quality of a service, in particular when deciding on the optimum vertical polar diagram of a transmitting aerial, the relative efficiency of the high- and low-angle modes in regard to both propagation and reception may play an important part. Therefore, if a comparison of such modes is attempted by means of a back-scatter display using transmission and reception from a representative transmitting aerial, allowances should be made on the one hand for the variation of back scatter from the ground with angle of incidence, and on the other hand for the nature of the vertical polar diagram of a typical receiving aerial that will be in use for the broadcasting or communication service under consideration.

In point-to-point communication one might consider the alternative of pulse-sounding the path by using a relatively low-power transmitter at the receiving end, with reception at the transmitting end. In this case a single transit and not a go-and-return path is involved, and representative aerials could be used at each terminal.

Mr. H. Page (*communicated*): The author has compared the observed back-scatter pattern with that expected on theoretical grounds, based on the measured parameters of the propagation path. The converse problem, that of interpreting the pattern in terms of the relative field-strength incident on the target area, seems to me to present insuperable difficulty. Since we are usually concerned with distances large compared with the height of the reflecting layer, there is little doubt about the range from which a particular echo is returned, but I do not see how we can accurately estimate the relative magnitude of the incident field. This depends on the degree of irregularity of the ground and also on the downcoming angle; the latter is not known, and even if it were, the author has to assume how it affects the ratio

* Hymans, A. J., and Lait, J.: 'Analysis of a Frequency-Modulated Continuous-Wave Ranging System', *Proceedings I.E.E.*, Paper No. 3264 E, July, 1960 (107 B, p. 365).

between the incident and scattered fields. The problem is even more complicated if signals are propagated to a given point by more than one mode.

I therefore suggest that the back-scatter display can only give information of a qualitative nature about the relative incident field-strength. If the author does not agree, will he describe precisely how he would set about interpreting a back-scatter display, and also say what accuracy of prediction of relative incident field-strength he considers could be achieved at various ranges?

THE AUTHOR'S REPLY TO THE ABOVE DISCUSSION

Mr. E. D. R. Shearman (*in reply*): I do not think very long-distance paths (greater than 5 000 km) are suitable for the technique. Echoes via three or more hops are frequently obtained, especially when much of the route is in darkness, but usually two-hop paths represent the limit for consistent reception. Satisfactory results have been obtained with east–west circuits, the observations revealing complex behaviour as the day/night boundary moves along the route.

I would stress differently from Mr. Jowett the factors which make back-scatter sounding less suitable for fixed than for broadcast and mobile services. Admittedly, experienced circuit operators develop sound judgment of the best frequency to use and the best times for frequency changes, often assisted by a watch on other transmissions over the route. Even so, an optimum-frequency measurement technique can help when conditions are unusual.

The important factor in fixed services is the much greater signal/noise margin. With high-gain aerials, high transmitter power, diversity reception and receivers with narrow bandwidths it is possible to work well below the m.u.f. in spite of D-layer attenuation. Back-scatter cannot compete with this performance, and consequently cannot assist in the choice of what may be operationally more convenient frequencies.

I agree with Mr. Phillips that for long-distance fixed circuits one-way sounding is a better solution. The Radio Research Station is carrying out experiments over a series of such circuits, using a sweep-frequency pulse transmitter at one terminal and a synchronized sweep-frequency receiver at the other.

The back-scatter technique seems most profitable for broadcasting services, and for communication with low-power mobile units such as aircraft. With the former it is difficult to get quick reports of the signals received by the listeners, and with the latter the low-power transmitters and low-gain aerials of the mobile units make the system performance comparable to that of back scatter.

The aircraft trials mentioned by Mr. Anslow gave some more evidence about the gap phenomenon. An aircraft transmitted at intervals during a transatlantic flight on a frequency high enough to give a gap between first and second hops. The field strength received in England fell to a low value when the aircraft flew between the ground areas from which one-hop and two-hop back-scatter was being received.

Mr. Hitchcock raises doubts about the value of the prediction of propagation modes from vertical-incidence data. However, once mode predictions are programmed for a computer it is not difficult to put in probable day-to-day variations in vertical-incidence parameters and so obtain advance information about the likely extent of the variation in mode structure.

The scheme Mr. Morris describes for pulsing one sideband of an independent-sideband transmitter is a very elegant solution to the previously intractable problem of using an operational transmitter for back-scatter sounding without interrupting the service it carries. A separate transmitter is needed, however, if it is desired to inspect the back-scatter echo patterns on alternative frequencies.

To discriminate against backward echoes, we have found it necessary when using rhombic aerials to use 3-wire construction, thus obtaining an adequate back-to-front ratio. Auroral echoes can be distinguished by their rapid fading which is uncorrelated between successive 25 c/s time-base sweeps. In comparison, ground-scatter fading rates are slow, being about $\frac{1}{2}$–2 c/s.

Mr. Merriman queries the restricted freedom in the choice of frequencies. The operational problem which we have been endeavouring to solve is how to use most effectively the frequencies already allocated to a service, not to suggest a large variety of other frequencies.

There should be no need for a back-scatter pulse transmission to cause adjacent-channel interference if Gaussian-shaped pulses of about 1 ms are used.

We have attempted to simplify the presentation of back-scatter echo information by using a chart-recorder in conjunction with a remotely controlled rapid frequency-change transmitter. A ship's echo sounder of the spinning-pen type is used, as the time delays for back-scatter correspond to those for sound waves returning from depths of about 100 ft. The chart gives a measure of the skip distance on four frequencies, the complete sequence taking about three minutes to record.

Mr. Phillips and Mr. Page stress the importance in back-scattering from the earth's surface, of the variation with the angle of incidence and the type of ground. The variation with type of ground appears surprisingly small; the only objects which have so far been identified are large mountain ranges. At present, the major unknown is the variation with the angle of incidence, which can only be measured satisfactorily from an aircraft. Mr. Page is therefore right in saying that only qualitative measurements of field strength can be made with present knowledge. However, the relative field laid down by two aerials of similar horizontal directivities can be measured.

It should be possible to measure the angle of elevation of back scatter from a particular range by the use of an auxiliary vertically sliding horizontal dipole aerial for reception. A null of calculable and adjustable elevation angle is then available. This gives rise to a trough in the echo pattern at the range corresponding to the elevation angle. By steering the null, the complete range/elevation-angle relationship can be found. The technique has recently been used in Norway for auroral-echo measurements.

The presence of two propagation modes to the same range gives rise to an ambiguity which does not appear to be resolvable.

I would like to make a plea for more information from operators of h.f. radio networks about the financial gain to them of any increase in circuit availability and continuity of service. This information would be very useful to us when developing propagation-evaluation techniques such as back-scatter, one-way sounding and mode prediction, or when assessing the need for more knowledge about path attenuation, atmospheric noise and the propagation of interfering signals.

621.397.13

The Institution of Electrical Engineers
Paper No. 3588
July 1961

B.B.C. TELEVISION 1939–60

A Review of Progress

By E. L. E. PAWLEY, O.B.E., M.Sc.(Eng.), Member.

(1) INTRODUCTION

A review of broadcasting and television was presented to The Institution by Sir Noel Ashbridge in 1938.[1] The further development of sound broadcasting in the United Kingdom during the years 1939–60 has recently been reviewed by the author.[19]

During the same period, and particularly since 1946, there has been continuous and remarkable progress in television. The present paper deals mainly with B.B.C. television, although reference has been made where appropriate to other major developments in this field both at home and abroad. It shows that the original engineering conception on which the B.B.C. television service was based when it opened in 1936 has proved fundamentally sound and capable of development to meet the much more complex requirements of the present day.

The B.B.C. Engineering Division is organized in such a way that, except on the operational side, sound and television are closely integrated. Developments common to both have been described in the companion review of progress on sound broadcasting.

(2) SITUATION IN 1939

The B.B.C. high-definition television service (which was the first regular public high-definition service in the world) was opened officially on the 2nd November, 1936, although programmes had already been transmitted experimentally during the Radio Exhibition at Olympia in August of that year.

The population covered by the one transmitting station in operation at Alexandra Palace before the war was about 12½ million, but the number of receivers in use by the summer of 1939—the beginning of the period now under review—was only about 23 000. This was a disappointingly small total after nearly three years of regular service, and certainly far below the number which would be needed to make the service financially self-supporting. The reasons for this slow rate of growth in the early years probably lie, for the most part, outside the scope of an engineering review, but it may be that a contributory cause was the public reaction to the earlier 30-line experiments, which were widely publicized but proved to be far too crude to have any entertainment value—except as a novelty. This may have given rise to a fairly widespread and persistent unwillingness to believe that practical television really had arrived this time. There is reason to believe, however, that the public response to the new service was on the point of increasing sharply at the very time when transmissions were stopped because of the imminence of war (on 1st September, 1939). Such were the inquiries from both trade and public at the 1939 Radiolympia Exhibition that the number of receivers might well have reached 80 000 by the end of the year if the transmissions could have continued.

By 1939 British television had reached a state of technical development that comprised all the essentials of a practical and economic television service. Neither the transmitted signal nor the performance of the television receivers of that day were as good as those of their present-day counterparts, but the quality of the received picture was nevertheless acceptable to the great majority of viewers. The 'live' studio cameras were insensitive and demanded extremely skilled operation of the electronic controls, but given adequate lighting and sufficiently skilled handling of the camera control units they produced remarkably good pictures. The 'live' outside-broadcast cameras were also less sensitive than those of to-day, but they could give good pictures out of doors in daylight even with a heavily overcast sky. Serviceable equipment was also available for transmitting pictures from films. There was no telerecording equipment, but this was not a fundamental barrier to a practical service. The price of a domestic receiver represented a considerably greater proportion of the average householder's earnings than it does to-day, but some of the sets were nevertheless within the means of a great many people.

Thus it is clear that while television engineering developments since 1939 have brought about a great increase in coverage, more numerous and more widely distributed studios, a wider range of potential sources of programme material, and larger picture screens, there have not been any revolutionary changes in the fundamental techniques. Those devised before 1936 have stood the test of time and still hold promise of further development, e.g. for colour television.

(2) POST-WAR RESUMPTION OF SERVICE

In September, 1943, the Government appointed a committee, under the chairmanship of Lord Hankey, to prepare plans for the reinstatement of the television service after the war.[2] The committee recommended that the service should be restarted in London on the same standards as those in operation before the war, rather than that the resumption should await the development of a new television system. It was recognized that research and development would be required before a service on a new standard could be made available to the public. It was considered important to restart the service as soon as possible so that this country might continue to hold a leading position in the television field and so that the highly specialized staff who had been working in television before the war, and had since been engaged on war work, might not become dispersed. At that time no alternative system giving appreciably higher definition had been developed in any European country. (Subsequently, however, standards using 625 and 819 lines were proposed for use in European countries and a meeting of the Television Study Group of the C.C.I.R. in July, 1950, attempted unsuccessfully to reach agreement on a common standard for Europe). The Hankey Committee also recommended that television should be extended, as soon as possible, to the most populous provincial centres on the same basis. The first post-war Government accepted these recommendations, and accordingly, after extensive overhauling and some detailed improvement of the London transmitter and studio equipment at Alexandra Palace, the service reopened on 6th June, 1946, on the same system (405 lines, 25 pictures per second with 2 : 1 interlace, positive modulation and a.m. sound) which had been in use before the war.

Mr. Pawley is with the British Broadcasting Corporation.

The fundamental question of standards was reviewed during the ensuing two years, and in 1948 it was decided by the Government that no change should be made. A new system using a number of scanning lines in excess of 405 would have offered some increase in vertical definition of the received picture, but the horizontal definition would still have been restricted by the need to limit the bandwidth of the transmissions in order to accommodate all the projected transmitting stations within the available frequency band. It has been found in practice that the B.B.C. standard of 405 lines is capable of giving excellent pictures in the home at lower cost than is possible with other standards, while at the same time making the best use of the limited frequency band that has so far been made available.

A minor alteration was made in April, 1950, when the aspect ratio of the transmitted picture was changed from 5 : 4 to 4 : 3 to bring it into conformity with standards adopted elsewhere. A convention on the British contribution to television was held by The Institution in 1952.[3]

(3) EXTENSION OF COVERAGE: BUILDING UP A NATION-WIDE SERVICE

Once the question of standards had been settled, it became possible to develop a plan for extending the television service beyond the area served by Alexandra Palace. A plan for nationwide coverage was prepared and approved by the Postmaster General. The plan envisaged five high-power transmitting stations serving, respectively, the London area and South-East England, the Midlands, the industrial North of England, Central Scotland, and South Wales with part of the West of England. The major centres of population lying between the areas covered by these five stations were to be covered by five medium-power stations.

These stations had to be accommodated within the band of frequencies (41–68 Mc/s) allocated to television broadcasting in 1947 by the Atlantic City Conference of the International Telecommunication Union. It was therefore decided that they should all use the vestigial-sideband system of transmission to reduce the total bandwidth required—apart from the London station at Alexandra Palace, which, being of much earlier design, had always used the double-sideband system. (This system was retained, until the Alexandra Palace station was superseded by Crystal Palace in 1956, because of the significant number of double-sideband and upper-sideband receivers still in use in the London area.) By this means five independent frequency channels were obtained.

The first of the high-power stations outside London was opened in December, 1949, at Sutton Coldfield, near Birmingham, and all four of the new high-power stations were completed by August, 1952. Construction of the medium-power stations was delayed and authorization by the Government to proceed with them was not received until 1953. In the early part of that year a special effort was made to increase the coverage to the maximum possible in time for the Coronation in June of Her Majesty Queen Elizabeth. Temporary stations were built with all speed at Glencairn near Belfast, at Pontop Pike, covering the heavily populated Tyneside area, and at Truleigh Hill, near Brighton.

Construction of the five permanent medium-power stations was started in July, 1953, and completed in June, 1956. Meanwhile a temporary low-power station had been brought into operation in the Isle of Man, and further stations had been authorized to cover East Anglia, the Channel Islands, part of the North of Scotland, the Londonderry area of Northern Ireland, North-West England, and West Wales. All these stations were completed before the end of 1957.

In March, 1956, the original Alexandra Palace transmitters were closed down and the service was transferred to a new transmitting station on the site of the Crystal Palace, in South London. This station was brought into service in a temporary condition pending completion of the permanent mast and aerial system and was completed by the end of 1957. The B.B.C. television service was then available to some 49½ million people—more than 98% of the population of the United Kingdom. Additional medium- or low-power stations have since been opened at Orkney, Thrumster (near Wick), Peterborough, Dover and Folkestone, bringing the coverage up to 98·8%.

The staffing of all these stations presented a major problem. It was solved mainly by the development by the B.B.C. of automatic devices, which made it possible to release staff from the sound services who were already trained in broadcasting engineering.

In June, 1959, the Postmaster General announced that the B.B.C. had been given permission to proceed with the construction of 14 (since increased to 15) satellite stations to be sited in areas where there was at present no service or where reception was unsatisfactory. In May, 1960, approval in principle was given to a further ten satellite stations. (See Appendix 20.2.)

(4) FREQUENCY PLANNING

(4.1) Bands Allocated to Television

The arranging of inter-Governmental conferences on the broad allocation of the radio-frequency spectrum among various services is the responsibility of the International Telecommunication Union, now a specialized agency of the United Nations. The frequency-band allocations currently in force were part of the radio regulations agreed at a conference at Atlantic City in 1947.[48] The range of frequencies then allocated was 10 kc/s–10·5 Gc/s. A revision of this allocation table was made at Geneva in 1959, the new agreement being operative from May, 1961, in which the range of frequencies has been extended to 40 Gc/s. The bands allocated to television and applicable to the United Kingdom, which do not greatly differ from those allocated at Atlantic City, are as follows:

Bane	*Frequency Range*	*Remarks*
Band I (V.H.F.)	41–68 Mc/s	Five channels for B.B.C. television service
Band III (V.H.F.)	174–216 Mc/s	Eight channels available in the United Kingdom, of which four are already used for the I.T.A. service.
Band IV (U.H.F.)	470–582 Mc/s	Not yet used in the United Kingdom.
Band V (U.H.F.)	606–960 Mc/s	790–960 Mc/s shared with other services.

(4.2) V.H.F. Bands

A realization of the importance of planning the v.h.f. broadcasting bands while development in many European countries was at an early stage led to the convening[4] of a European Regional Conference at Stockholm in May and June, 1952. Some 30 European countries were represented at the conference, many of whom had at that time no immediate plans for developing television or v.h.f. sound broadcasting services but who wished to see possible future services catered for in an agreed European plan.

The conference produced plans for the use of v.h.f. Bands, I, II and III, and, realizing that these plans could not be too rigid owing to the uncertainty of future developments in most of the countries represented, suggested that a further conference be held in 1957. (It will be held in May, 1961.)

In planning national coverage for the B.B.C. television service,

it was anticipated that frequencies in Band III would be made available to the Corporation in accordance with the Stockholm plan. No such frequencies have, however, yet been allocated and it has therefore been necessary to accommodate the 23 existing stations and the 25 proposed 'satellites' in the five channels of Band I.

This has required much planning effort in order to reduce as far as possible co-channel interference between existing and proposed stations. The choice of channel is in some cases difficult, but once it has been made, co-channel interference can be minimized by the following means:

(*a*) The use of vertical polarization at certain stations and horizontal polarization at others, resulting in a reduction of at least 10 dB in interference between a pair of stations.[38]

(*b*) Shaping the horizontal radiation patterns of the transmitting aerials where appropriate. (This has also to be done in the case of certain stations which operate on frequencies close to those of nearby Continental stations, to reduce the power radiated in the direction of the Continental station below a specified level.)

(*c*) Off-setting the carrier frequencies of neighbouring pairs of co-channel stations, normally by 6·75 kc/s (two-thirds of the line scanning frequency) in the case of the vision carriers and 20 kc/s for the sound carriers.

It has been found possible to keep co-channel interference within reasonable bounds, but its occurrence in certain areas for a proportion of the time and particularly at certain seasons of the year is inevitable and gives rise to numerous complaints from the public.

An experiment recently concluded by the B.B.C. has shown that a further reduction in co-channel interference of at least 6 dB can be achieved by the use of precision-frequency offset of the vision carriers. In this experiment, carried out on Channel 2, the frequency of the vision carrier of the high-power transmitter at Holme Moss was maintained within very close limits by a highly-stable crystal drive, which was itself manually corrected when necessary by reference to the standard frequency transmissions from Droitwich on 200 kc/s. At North Hessary Tor—a medium-power station also operating on Channel 2—a vision carrier frequency was generated differing from that of Holme Moss by two-thirds of the instantaneous line frequency, ± 2 c/s.

The Band I frequencies are also seriously affected by interference from transmitting stations outside the United Kingdom. In winter, during the years of maximum sunspot activity, strong interference has been experienced from police and business radio stations in the United States. Their signals are propagated at such times via the ionosphere and affect mainly B.B.C. stations on Channel 1. Interference from television and other transmitting stations in Europe has also increased in recent years with the opening of additional stations using frequencies in Band I. Propagation in such cases is by the tropospheric or sporadic-E modes.

During the past two years, the opening of high-power stations using the forward-scatter system of communication has constituted a very serious threat to television reception in Band I. The frequencies used by some of these stations lie near the standard vision intermediate frequency adopted by British television receiver manufacturers (34–38 Mc/s), and it is sometimes impossible to eliminate the interference entirely, even with very efficient filtering at the receiver, when it is tuned to Channel 1.

The recent appearance of 'scatter' stations in Channel 1 itself is even more ominous, although some protection against this form of interference and also against interference from scatter services operating near the intermediate frequency of receivers should be afforded by the agreement reached at the Administrative Radio Conference held at Geneva in 1959.

If channels in Band III were made available to the B.B.C., they would enable further improvements to be made in the coverage and would also make it possible for the B.B.C. television programmes for a large part of Wales to be separated at times from those transmitted over the main network. Such channels might also afford a means of escape from the serious interference that affects Band I (and particularly Channel 1).

(4.3) U.H.F. Bands

The two further bands of frequencies (Bands IV and V) allocated internationally for television have so far been used only on a very limited scale in Europe. The B.B.C. has made a long-term study of propagation conditions in these bands, which is still continuing. The information obtained in this way is, however, insufficient to determine fully the suitability or otherwise of Bands IV and V for television broadcasting in the United Kingdom, and at the request of the Television Advisory Committee the B.B.C. embarked, in collaboration with other organizations, on a more ambitious series of experiments using a high-power transmitter radiating full television signals. High-power vision and sound transmitters were therefore installed at the Crystal Palace television station in November, 1957, together with a high-gain aerial mounted on a 40 ft pole at the top of the 668 ft mast. The transmitters were operated on frequencies around 650 Mc/s, and in the initial series of tests, which concluded at the end of March, 1958, they radiated television signals on 405 lines, enabling a direct comparison to be made with the B.B.C.'s normal Band I transmissions. A further series of tests using television signals on the C.C.I.R. standard of 625 lines began in May, 1958, so that comparisons could be made between the results obtainable from the two systems in these bands. A comprehensive report on these field trials was published[39] by the B.B.C. in May, 1960; a report by the Television Advisory Committee embodying their recommendations for the future technical development of television in the United Kingdom was published[40] in June, 1960. The decision on the future use of these bands in the United Kingdom rests with the Postmaster General, who will have available to him the advice of the Television Advisory Committee.

(5) VISION AND SOUND TRANSMITTERS

(5.1) General Design Considerations

At all stages in the development of the B.B.C. television service, the industry has been able to produce transmitters capable of providing the bandwidth and effective radiated power required. Transmitter design, in fact, has never been a limiting factor in the growth of the service. Nevertheless, it is particularly important that the capital and running costs of the transmitters should be low, as the studio and programme costs are necessarily so high. Much attention has therefore been devoted to improving transmitter performance, particularly in respect of efficiency and reliability and in providing more compact transmitters which need smaller and less expensive buildings. The question of losses in the feeders connecting transmitters and aerials has also received close attention.[27] Many of the developments mentioned in connection with sound broadcasting stations apply also to television stations. In addition, the wide bandwidths required for television and the fact that both sound and vision components have to be transmitted have posed special problems. The respective merits of low- and high-level modulation have been discussed elsewhere,[28] and there is little doubt that, if tetrodes are available as grid-modulated r.f. amplifiers, the choice is, wherever possible, in favour of high-level modulation. In the first B.B.C. television stations omnidirectional aerials were used, but the need to provide a large number of stations on a

limited number of channels made it necessary to design directional aerials, so as to reduce the interference between stations on the same, or adjacent, channels. The masts have had to be designed in many cases to accommodate high-gain television aerials, either vertically or horizontally polarized, in addition to aerials for v.h.f. sound transmissions.

At the first B.B.C. television station at Alexandra Palace the sound and vision components were transmitted from separate aerials, but in later stations the two components have been combined and fed to a common aerial; in some cases the transmitters are installed in pairs, one of which feeds into one half of the aerial and the other transmitter, correctly phased, into the other half of the aerial, so that in case of breakdown, transmission can continue uninterrupted on reduced power. Several improvements have been made in the design of the combining circuits, which are now located in the transmitter building instead of at the top of the mast. At first tuned rejectors were used, but in later designs a constant-impedance bridge circuit has been employed so as to present the correct load to the transmitter over the full video bandwidth. Because of the high gain of the aerials in the horizontal plane, it has sometimes been necessary to take special steps to fill in areas of weak or distorted reception very close to the transmitting station.[29, 30]

The basis of the design of a high-power television transmitting station is to provide a field strength high enough to give adequate signal/noise ratio at the maximum range compatible with reasonable freedom from fading and from co-channel interference under average reception conditions. With the limited knowledge of v.h.f. propagation available when the London television station was planned in 1935, it was thought that the fading-free range would be about 30 miles, and on this assumption the Alexandra Palace site appeared to give the best population coverage. Moreover, part of the building was available for conversion to studios.

(5.2) First London Station

The Alexandra Palace station had a nominal effective radiated power of 34 kW on vision, with a 17 kW transmitter and a 2-tier aerial giving 3 dB gain. High-level modulation was fed to the grids of the CAT9 final r.f. amplifiers. The high-voltage supply was obtained by rectifying the output of a 500 c/s motor-alternator. Separate aerials and feeders were used for the vision and 3 kW sound transmitters.

(5.3) Post-War Stations

Experience with the Alexandra Palace transmitter had shown that the average fading-free range was considerably greater than the 30 miles originally assumed, and that an acceptable standard of reception was possible with field strengths down to 0·1 mV/m in reasonably interference-free conditions. However, in order to override interference as much as possible, it was desirable to provide some 0·25 mV/m at the limit of the fading-free zone. These factors warranted the use of a considerably higher e.r.p. than the 34 kW of the Alexandra Palace, and in the meantime valves had been developed which would make this practicable.

The first stage in the post-war extension of the coverage was the commissioning of four new high-power stations with vision transmitter powers of 50 kW and e.r.p.s of 100 kW, which would, together with the Alexandra Palace station, bring the B.B.C. television coverage up to 80% of the population of the United Kingdom.

Each of these stations was also designed to serve in due course as a high-power v.h.f. f.m. transmitting station, to radiate three, and in one case four, sound programmes. The sound transmitters were to be installed in the same buildings as the television equipment and to operate on a semi-attended basis with automatic alarm devices to give an indication in the adjoining television control room—where engineers would be on duty—when anything was amiss. A cylindrical slot aerial, some 90 ft in overall length and radiating horizontally polarized signals was to be mounted on the mast immediately below the television radiator, to carry the sound programmes.

(5.4) Sutton Coldfield

The station at Sutton Coldfield, near Birmingham, which was brought into service in December, 1949, was the first of these high-power television stations.[31] This was, like Alexandra Palace, high-level modulated at the grids of the final r.f. amplifiers, but the CAT 21 water-cooled triodes used in this stage gave a peak-white transmitter power of 45 kW. The power was later increased to 50 kW peak white when BW 165 thoriated filament valves became available. When Sutton Coldfield was opened there were no standby transmitters. A 5 kW vision transmitter and 1·25 kW sound transmitter were added in April 1953.

(5.5) Holme Moss

The next high-power station to be completed was at Holme Moss, between Manchester and Huddersfield, and it came into service[32] in October, 1951. It is generally similar to the Sutton Coldfield station, but medium-power standby vision and sound transmitters were included from the start. Reserve aerials were originally mounted on a separate 150 ft mast, with separate vision and sound aerials. A single reserve aerial has since been mounted at the 470 ft level on the main mast, and is fed with the combined vision and sound transmitter outputs through a separate feeder. Owing to the extremely exposed site, which is 1 700 ft above sea-level, the brick building has been covered in an outer sheath of stone, and the possibility of staff being snowbound on the site for several days at a time has been provided for. Like its two predecessors, the main vision transmitter uses high-level grid modulation.

(5.6) Kirk o'Shotts and Wenvoe

The two remaining high-power stations were at Kirk o'Shotts, serving the more densely populated parts of Scotland, and at Wenvoe, near Cardiff, serving a large part of South Wales and South-West England.[33, 34] These began regular service in March and August, 1952, respectively. The buildings and equipment of both stations were similar, and both broke away from previous practice by using low-level modulation in the main vision transmitters. This was made possible by the fact that valves and circuits for wide-band r.f. amplifiers had become available.

An interesting feature of these stations is the use of a suspended locked coil rope forming the inner conductor of the vertical feeders on the mast.

(5.7) Medium-Power Stations

The next stage was to increase the coverage to about 97% of the population by seven medium-power stations with vision transmitter powers of 5 kW and e.r.p.s of about 12 kW. These stations had perforce to use the same channels as the five high-power stations, and co-channel interference was minimized by the means described in Section 4.2. All these stations were in operation by November, 1956.

Like the regional high-power stations, the medium-power stations were designed to incorporate multi-programme high-power v.h.f. sound broadcasting stations sharing the same buildings and aerial mast.

(5.8) Crystal Palace Transmitters

While the construction of the regional high-power and medium-power stations was proceeding, a new high-power station for South-East England, with an e.r.p. of 200 kW, as compared with the 34 kW of Alexandra Palace, was being planned.

With the increased range planned for the new station a more southerly site would give improved coverage in Kent, Surrey and Sussex without appreciable deterioration of the service to the north and north-east. There would, of course, be a considerable decrease in field strength in the immediate neighbourhood of Alexandra Palace, but this area would still receive an adequate service.

After considering a number of sites, it was decided that the best position for the new station would be on the site of the Crystal Palace. Field-strength measurements, using a transmitting aerial supported by a balloon, showed that an additional two million people would be covered within the area enclosed by the 0·1 mV/m contour, as compared with the corresponding contour for the Alexandra Palace station. A 2-acre site in the Crystal Palace grounds was accordingly negotiated with the L.C.C. Completion of the new station was delayed by factors outside the B.B.C.'s control, but it was brought into service in a temporary condition on 28th March, 1956. With a temporary mast and aerial, an e.r.p. of 60 kW was obtained. The Alexandra Palace transmitters were then finally closed down. On 10th September, 1956, a temporary aerial was brought into service on the main tower, giving an e.r.p. of 120 kW, and on 18th December, 1957 the final 8-tier aerial giving the full e.r.p. of nearly 200 kW was commissioned.[35, 36, 37]

Space is provided for increasing the power of the vision and sound transmitters to 50 and 12 kW, respectively, if and when the internationally-authorized e.r.p. of 500 kW should be used. With the same possible end in view, the aerials, feeders and aerial switching are designed to handle a power of 100 kW.

(5.9) Low-Power and Satellite Stations

A number of stations of lower e.r.p. were subsequently built, and the B.B.C. is now implementing stages I and II of its scheme for low-power satellite transmitters. A new technique has been developed by B.B.C. engineers for satellite stations using equipment known as a 'translator'. This receives the sound and vision signals by radio from another B.B.C. transmitting station and retransmits them on a different channel. The carrier frequencies of the sound and vision signals are retransmitted without demodulation, which simplifies the installation and increases its reliability. It has therefore been possible to design these stations to operate without staff in attendance. The first of these stations was put into experimental service at Folkestone in July, 1958.

A similar idea has been widely used on the Continent for serving small areas screened by high ground from the main transmitters. Translators for the British television standards are more difficult to design than their Continental counterparts, as the British use of amplitude-modulated sound makes it necessary to amplify the sound and vision signals separately to avoid cross-modulation, and to provide separate a.g.c. systems.

(6) TELEVISION STUDIOS

(6.1) London Studios

Before the war, and from the resumption of the service until 1950, all studio programmes came from two studios at Alexandra Palace. With a view to extending the service after the war, a 13½-acre site at the White City in Wood Lane, near Shepherds Bush in West London, was acquired in 1949 for development as a multi-studio television centre. Since it was clear, however, that a number of years must elapse before any studios on this site would be available for programmes, steps were taken to provide interim studio accommodation. The studios at Lime Grove (previously used for films) were acquired in 1950, and four large general-purpose television studios, together with a small presentation studio, were brought into use in this building at intervals during 1950–53.

For 'light entertainment' programmes requiring a large studio audience, the old Shepherds Bush Empire theatre was acquired in 1953.

One of the Alexandra Palace studios was used from July, 1954, for the television news and newsreel service, which has its own independent camera, telecine and telerecording facilities.

In 1954 the B.B.C. acquired another film studio building known as the Riverside Studios, on the Thames, near Hammersmith Bridge. This provided the Corporation with two additional studios and made an important addition to the London studio facilities. Although these had been built, like the Lime Grove studios, for film-making purposes, their conversion for television was less hurried than the Lime Grove adaptations. It was also possible to install more advanced technical equipment than in any of the earlier studios, especially as regards the methods of control of the production lighting.[5,6]

Small single-camera interview studios have been equipped near Broadcasting House, London, at London Airport, and near the Houses of Parliament. The latter studio is operated on an 'unattended' basis, the television equipment and lighting being switched on from the Television News Headquarters at Alexandra Palace as required. The studio near Broadcasting House is equipped with a remotely controlled camera developed by B.B.C. engineers. The camera controls (pan, tilt, focus, zoom and iris) can be operated from a control panel at Alexandra Palace some six miles away.

Planning of the new Television Centre began in 1949. Its construction started in 1951, and the first of the buildings to be completed was the scenery block in 1954. This was brought into use for the construction and storage of scenery and properties to supply the various existing London studios. The restaurant block was built next and was used temporarily from October, 1955, to provide additional offices and rehearsal rooms. It was completed by June, 1960, when the first of the studios and the main block were opened. The plan provides for seven production studios, of which four are being completed and equipped initially. The Centre is now the hub of the whole B.B.C. television network and contains, apart from the various technical areas necessary for the operation of the studios, a central apparatus room and presentation suite, comprehensive facilities for the transmission of films, and telerecording equipment using film and magnetic tape.

The second of the studios was brought into operational use in January, 1961, and two others have since been opened.

(6.2) Regional Studios

Television studios capable of handling major programmes have been established at Birmingham, Bristol, Cardiff, Glasgow, Manchester and Belfast.

There are also smaller studios for news and interviews at Birmingham, Bristol, Cardiff, Glasgow, Manchester, Norwich, Newcastle, Belfast and Southampton.

(7) CAMERAS AND CAMERA TUBES

When the B.B.C. television service reopened in 1946, the Emitron (iconoscope) and Super-Emitron (image iconoscope) were the only cameras in use.[7, 8]

In November, 1947, two experimental cameras, using the new cathode-potential stabilized Emitron tube, were used in the outside broadcast of the wedding of H.R.H. Princess Elizabeth and H.R.H. the Duke of Edinburgh. This camera tube, a development from the orthicon, was first produced in 1939, but was not then used in service owing to the war.[10] Like the iconoscope, it uses a single electrode as both photocathode and target. This electrode is stabilized at the potential of the gun cathode. Although there is no electron multiplication before scanning takes place, the tube utilizes the photo-emission more efficiently than the iconoscope, and its sensitivity is considerably greater than that of either of the other two tubes.

The tube had, however, an inherent disadvantage: the mosaic became unstable if excess light fell on any part of it, even for a brief period. Since it is difficult to avoid occasional flashes of excess high-light brilliance, especially in outside broadcasts, this was an important operational limitation. Nevertheless the commercial version of the c.p.s. camera found some application in outside broadcasts until the even more sensitive image orthicon became available. In the studios it is easier to control the maximum high-light brilliance and so avoid instability, and full advantage can then be taken of other features of the tube. These are the linear transfer characteristic and the fact that the black level is fixed. When used with external gamma correction, these features enable pictures with extremely good tonal gradation to be produced. By the beginning of 1950, when Studio D in the newly-acquired Lime Grove film-studio building was being prepared for television operations, the image orthicon had become accepted as the best camera tube for outside broadcasts, and some c.p.s. camera channels designed for outside broadcasts were released for studio service.

In December, 1950, Studio G at Lime Grove was brought into service using Photicon-type miniature image-iconoscope outside-broadcast camera channels.

These particular camera channels were better suited for this studio than the c.p.s. channels would have been, as the studio was mainly used for light-entertainment programmes in which a moving spotlight sometimes produced high contrast ratios which would have been difficult to cope with on the c.p.s. channels.

In a modified type of c.p.s. tube, introduced in 1956, the tendency to instability in the mosaic was largely overcome by an additional mesh on the scanning side of the target. The early specimens of this tube had somewhat low sensitivity, but a new tri-alkali mosaic, introduced towards the end of 1957, increased the average sensitivity about three times. It has now become possible to standardize on a lighting level of 75 ft-candles in studios, and under these conditions the c.p.s. tube produces very good pictures.

The image-orthicon camera tube (first described in 1946) came into use in Great Britain in 1949.[11, 12] In this tube the photocathode and target are separate, as in the image iconoscope, but the target is stabilized at gun-cathode potential, as in the orthicon. Instability of the type encountered in the c.p.s. tube is prevented by a mesh very close to the target on the photocathode side, and the sensitivity is greatly enhanced by directing the return scanning beam into an electron multiplier incorporated in the tube. The original 3 in image orthicon exhibited a number of drawbacks of which the main ones were noise, black halo around high lights, poor definition and inability to produce true edge transitions. These faults tended to offset the advantage of greater sensitivity which the tube possessed over its contemporaries, but the very high sensitivity (up to 1 000 times that of the iconoscope) soon established it as the preferred tube for outside-broadcast use.

Cameras with image-orthicon tubes were introduced for studio use in November, 1953, when Studio E at Lime Grove was brought into service. These camera tubes had 3 in targets. Further development of this tube by the English Electric Valve Co. led to the use of a $4\frac{1}{2}$ in target, which provides a means of increasing the storage capacity and hence the signal/noise ratio without reducing the life expectancy of the tube. Other improvements in the design of this tube, including the separation of the field mesh voltage supply from that of the focusing electrode, have allowed it, when carefully set up, to produce pictures of exceptional quality; it has now largely replaced the 3 in tube both for outside broadcasts and for studio use and will be used in all cameras at the B.B.C. Television Centre.[13]

The image-orthicon cameras used at the Television Centre have been designed to a B.B.C. specification which requires that only two operational controls—one varying the light input by means of a remotely controlled iris and the other the picture black level—need be extended to the vision-control supervisor's position. The specification requires that the remaining controls numbering about 30 can be preset in advance. This degree of stability has been attained by the manufacturers, with the result that it is now possible for one operator to control electrically all the cameras in a studio, normally four.

The photoconductive (vidicon) camera tube was first used experimentally in Great Britain on an outside broadcast in February, 1953, and has since been used to a limited extent in live broadcasts, although tubes of this type are at present more widely used in telecine equipment and for industrial television. In this tube the optical image is formed directly on the target and the conductivity of each portion of the target varies with the light falling on it. The resolution is equivalent to a drop of about 12 dB at 3 Mc/s, the sensitivity varies at different portions of the target, and the picture is liable to shading effects. There is also a tendency to smearing on moving objects unless the scene is very brightly illuminated. However, the camera has the advantage of compactness and relatively low operating cost and it can give satisfactory pictures from simple subjects in small brightly-lit studios. For these reasons it has been used for outside broadcasts under cramped conditions, such as the interior of a submarine, and it is also used in interview studios.

The vidicon tube is sufficiently stable in operation to enable it to be switched on remotely and worked without any technician in attendance. This has made it possible to operate an 'unattended' studio near the Houses of Parliament. The camera channel and production lighting in this studio are switched on from the news studio at Alexandra Palace, and the parliamentary correspondent has only to switch on the microphone amplifier when he enters the studio. The same type of tube has also been incorporated in a radio camera which enables a single person to move about freely with the camera in his hand, and transmit pictures to a fixed receiver from a radio transmitter on his back. This camera channel is described in Section 13. A remotely-controlled version of the vidicon camera is also in experimental use at the All Souls interview studio, as already mentioned.

(7.1) Lenses for Television Cameras

A notable improvement in picture quality has resulted from investigations by the B.B.C. Engineering Division into the performance required from lenses on television cameras, which is very different from that required of lenses used in photography. This difference results from the bandwidth limitation in television. As a result it is convenient to express the performance of a lens in terms of its spatial frequency-response characteristics.[14] A new method of recording the spatial frequency-response characteristic of a lens automatically has been described by Hacking.[15] Equipment has been developed by the B.B.C. which enables rapid measurements to be made of

he performance of a camera lens by the operations and main-
enance staff.[16]

Zoom lenses have been used in conjunction with television :ameras for a number of years. As early as 1951 an improved :oom lens with a range of variation in focal length as great as i : 1 was introduced, while in 1956 there was a further notable mprovement when a lens having alternative ranges of 4–20 in ›r 8–40 in was introduced. A further development during the ›ast year has resulted in a lens with an overall variation in focal ength of 4–40 in in four ranges. In this case the required range :an be selected by a simple 'optical' switch which alters the elationship of the moving elements of the lens. This operation nay be carried out during transmission, the original picture lissolving smoothly into the new one; formerly it was necessary o dismantle and reposition the lens in order to change its ›perating range.

Zoom lenses for use in studios have recently been produced. ›reviously the size and weight, together with the minimum listance at which objects could be focused, had restricted heir use to outside broadcasts. These improved lenses have ›een developed and produced by British manufacturers in co- ›peration with the B.B.C.

The most recently developed zoom lens has an extended back- ocus distance of about 11 in, making it possible to mount the ens within a camera body so that it is alongside the tube yoke. 'his is made possible by the use of mirrors to fold the long ›ptical path from the back element of the lens to the tube face. 'he arrangement enables a more compact assembly to be used, nd cuts down the total length of the assembly normally required vith the 'in-line' type of zoom. The further addition of a wide- ngle adaptor enables such a camera to cover focal lengths of –40 in in four ranges, any of which may be selected at will by ontrols available to the camera operator.

(7.2) Camera Mountings

The design of the mountings for television cameras has dvanced considerably since the early days of the television ervice. Since the requirements differ from those for either still r film-studio cameras, it has largely been necessary to develop quipment specially for television purposes.

In the very early days the cameras were mounted on wooden ripods which stood on plywood bases mounted on castors. Iowever, the need for two main modes of camera transport oon became apparent. One of these was 'tracking', i.e. moving camera towards or away from the subject while it is transmitting picture. The other was 'crabbing', which is the moving of a amera in any direction across the studio floor while the camera emains pointed in the same direction.

For tracking, some form of steerable vehicle, usually with four vheels, was indicated, and in one early attempt to produce such vehicle, small motor-car wheels were used. Later the Path- nder-type dolly was used; this was a four-wheeled solid-tyred rolley, with the camera and cameraman's seat mounted on a hort boom which could be raised or lowered by a handwheel.)ollies have since been developed in which the panning and aising or lowering of the camera are power-operated under the ontrol of the cameraman, while the power-driven movement of ne whole dolly, and also the steering, are controlled by the dolly perator.

The other main type of camera movement, namely 'crabbing', generally provided by a three-wheeled pedestal in which each heel can pivot through a full 360°. If all three wheels are rned parallel to one another and linked mechanically so that ey remain parallel as they are pivoted, the whole pedestal can e made to roll in any direction across the studio floor, without tering the direction in which the assembly is pointing.

Another widely-used type of camera mounting is the camera crane, in which a four-wheeled trolley carries a double-ended boom resembling a seesaw, with the camera and cameraman on one end of the boom and a counterweight on the other end. This enables the camera to be moved laterally while on transmission, and to be raised or lowered farther than the fully power-operated type of dolly allows. It thus permits a wider range of movement than the dolly, but it occupies more space on the studio floor, and, since the swinging and tilting of the booms are not power-operated, it requires more personnel.

(8) FILM TRANSMISSION EQUIPMENT

The choice of a picture frequency of 25 per second for British television was dictated by the need for synchronizing the transmitting and receiving equipment with the 50 c/s supply mains. Sound films shot at 24 pictures/sec can be run at 25 pictures/sec without any serious adverse effects on either picture or sound.

Until 1949, the normal method of televising films was to project the film on to the photocathode of a television camera. As a normal intermittent film projector would produce flicker effects because the film pull-down time extends beyond the field blanking period, a Mechau continuous-motion projector was used. This machine projects a moving picture of steady brilliance without any blacking out. The picture quality was limited, however, by the use of a contemporary camera tube, and there was a further deterioration caused by mechanical and optical imperfections in the projector.

The commissioning of the first twin-lens flying-spot telecine machines in 1949 raised the standard of film transmission to that of the best 'live' pictures. In these and other flying-spot machines, an optical image of a synchronized but unmodulated raster of high brilliance is focused on the film, so that the light is modulated by the density of each portion of the film image and is transformed into a video signal by a photomultiplier. In the twin-lens system the picture gate is the size of a normal film frame, but the flying-spot raster has an aspect ratio of $4 : 1\frac{1}{2}$ instead of the normal $4 : 3$. The double-lens system focuses two images of the raster (one above the other) on the film in the gate. The film moves continuously downwards at the rate of one frame height in 1/25 sec, and the field scan moves in the opposite direction to the film travel. Thus the vertical movement of the scanning spot on the film itself is the resultant of the raster frame scan and the film motion. As each film frame passes through the gate, it is twice scanned from top to bottom by the two successive images of the raster, one corresponding to the even-line field and the other to the odd-line field. A mechanical shutter interrupts each raster image alternately to ensure that the film frame is scanned once only by each field.

The twin-lens machines can only operate when the film is synchronized with the television field frequency, but there is an operational need for a telecine system which will work at any film speed from zero upwards. This arises when, for example, film inserts are included in live programmes and it becomes desirable to have the minimum time delay between cueing and fading up of the telecine picture. The need is supplied by telecine channels which use the old Mechau movement with a flying-spot scanning system, instead of a storage camera tube. When used in this way, the Mechau movement causes the image of the flying-spot raster to remain stationary relative to the film, regardless of the rate at which the film is moving. Another type of non-synchronous film scanning equipment has been produced, using a rotating glass polygon to provide optical compensation.

Recently a number of telecine channels, using photoconductive camera tubes have come into use, especially in the Regions. In these machines, a standard 35 or 16 mm intermittent film pro-

jector focuses an image on the target of the photoconductive tube, and the relatively long storage time of this type of tube bridges over the time intervals when the optical image is blacked out for the film pull-down. With these channels the picture quality is limited by the inherent characteristics of the photoconductive tube (mentioned in Section 7), but the use of standard intermittent projector movements lowers the capital cost considerably, while the long life of a photoconductive tube makes for low maintenance cost.

Facilities for reproducing sound from magnetic track (either on the same film as the picture or on separate sprocketed film) are being added to all telecine equipment. This offers a higher standard of quality than an optical sound track can achieve after the several dubbings which it must usually undergo before reaching the final positive film.

(9) TELERECORDING

The use of recorded television signals, or telerecordings, has become an important part of television operations.

At present the basic method of making telerecordings on film is to make a film of the television picture with a motion-picture camera, but there is a difficulty which must always be circumvented by one means or another. A normal intermittent-motion film camera must interrupt its exposure for at least 10 ms during each film pull-down. This interruption must not be allowed to overlap the active scanning period of the television signal, and it should therefore be completed during the field blanking period. This period, however, only lasts about 1·4 ms in a system with 50 fields/sec. The means devised to overcome this difficulty are now described.

The B.B.C. began to make regular use of film telerecordings in November, 1949.* The method used at that time was to display the picture on a high-quality monitor and make a film of it by using a Mechau continuous projector adapted as a film camera. The optical compensator in this machine caused an image of the displayed picture to keep pace with the continuously-moving film for the duration of a full frame, without interruption for pull-down. A complete picture (odd and even fields) could therefore be recorded, but the action of the optical compensator was imperfect and caused some deterioration of the recorded image.

There was an urgent need to have at least one improved telerecording system available in time for the Coronation in June, 1953. To meet this need, the performance of the Mechau system was improved by both optical and electronic modification, and the so-called suppressed-frame system was also developed.[17] In this latter system an intermittent film camera was used, and the displayed picture was blacked out electronically on each alternate field to allow time for film pull-down. By this means, only half the active lines were recorded but the full horizontal resolution was retained. In spite of the halving of vertical resolution, this system gave very acceptable results and remained in use until 1957, when the equipment was modified to work on the stored field system, in which both fields are scanned and recorded.[18] Since part of each alternate field must be scanned while the film camera is shuttered for pull-down, this system relies on the afterglow of the display-tube phosphor to store this field until the shutter reopens. During the scanning of the stored field, the output of the video amplifier is automatically increased to compensate for the decay during storage, and this ensures that both fields have equal brilliance when exposed together. The stored-field system gives a standard of quality which is among the highest of any of the methods available, at the time of writing, for producing a telerecording which can b[e] projected optically and reproduced by any television system.

Another method of film telerecording, which has potentially [a] very high standard of quality, is by means of an intermitte[nt] film traction mechanism which can complete, or very near[ly] complete, the pull-down during the field blanking period. Th[is] requires very rapid acceleration of the moving parts of t[he] mechanism; the principal difficulty is not, as might be suppose[d,] the tearing of the film, but the avoidance of excessively rapi[d] wear in the mechanism itself. A 16 mm 'fast-pull-down' tel[e]recording channel has been in use at Lime Grove for some tim[e] and gives picture quality as high as is likely to be obtainab[le] from 16 mm film. A 35 mm fast-pull-down channel has sin[ce] been added.*

The recording of vision signals on magnetic tape offers impo[r]tant advantages; in particular, it provides the facility of imm[e]diate playback and saves the cost of the photographic film a[nd] its processing. Magnetic tape can, of course, by used aga[in] once the recorded programme is no longer required.

In 1956, therefore, the B.B.C. began to design equipment fo[r] this purpose which became known as Vera (vision electron[ic] recording apparatus). This equipment was first demonstrate[d] publicly on 8th April, 1958, and was used in the programm[e] *Panorama* in the following week. The vision signals a[re] recorded longitudinally on a ½ in wide magnetic recording tap[e.] The tape speed is 200 in/sec. A band-splitting system is used t[o] divide the vision signals into two bands of approximate[ly] 0–100 kc/s and 100 kc/s–3 Mc/s. The 0–100 kc/s band is mad[e] to frequency-modulate a carrier which is recorded on one trac[k.] The low-frequency content of the vision signal is thereby trans[-] formed to a frequency band corresponding to shorter wave[-] lengths so that the low-frequency and long-wavelength difficulti[es] inherent in the conventional magnetic recording system a[re] avoided. The higher vision band, from 100 kc/s upwards, [is] recorded simultaneously on a second track in a convention[al] manner, and the sound signals are recorded on a third trac[k] using an f.m. carrier as with the lower vision band.

Vera is an interesting development with possible application[s] in many fields, but for television broadcasting it has been supe[r]seded by equipment developed in the United States. This us[es] a magnetic tape 2 in wide running at a tape speed of abo[ut] 15 in/sec. The high recording head-to-tape speed essential f[or] the recording of the high-frequency components of the visio[n] signal is achieved by the use of four recording heads mounte[d] at the outer circumference of a rotating drum, with their ga[ps] parallel to the drum axis, and the vision signals are recorde[d] transversely rather than longitudinally on the tape. The 2 i[n] diameter drum rotates at 15 000 r.p.m., giving a relative head-t[o-]tape speed of about 1 600 in/s.

The telerecording of vision signals on magnetic tape giv[es] pictures of excellent quality but has the disadvantage, compar[ed] with film, that the recording is tied to the standards of a pa[r]ticular television system. Some form of standards convert[er] must therefore be used if the recording is to be used by a tele[-] vision service using different standards. Recordings on magne[tic] tape are less readily edited than those on film.

(10) TELEVISION SOUND

When the B.B.C. 405-line service was started, f.m. sou[nd] broadcasting was in the experimental stage. As a result t[he] service started, and has since become established, with an a.[m.] sound channel. Most other television services started at a la[ter] date and adopted f.m. sound. Although frequency modulati[on]

* British Patent No. 648924.

* British Patent No. 859030.
British Patent No. 714736.

ffers greater resistance to certain types of impulsive interference, iis advantage is not as important for television sound as it is or 'sound only' broadcasting. In a receiver for the British tandards, with limiters of average efficiency in both sound and ision circuits, it is almost always the picture that sets the limit o the degree of interference that can be tolerated.

(10.1) Microphones

The problems of microphone technique in television are in ome ways more difficult than in sound broadcasting.[19] There re two fundamental factors in television studio production hich combine to produce much of this added difficulty. The rst is that the action is commonly switched between sets on ifferent parts of the studio floor during the course of a programme, and the second is that, with few exceptions, the microphone must not be allowed to appear in the transmitted picture. or these reasons, it is not possible to use the accepted sound udio technique, in which the microphone is the central feature the studio and the artists are positioned around it in such a ay as to obtain the required aural perspective. As the action oves from one set to another, either the microphone must ove with the artists or fixed microphones in different positions ust be brought into use. At the same time, the need to keep e microphones out of 'shot' rules out many positions and types f mounting which would be perfectly satisfactory in a sound tudio.

In the early programmes from Studio A at Alexandra Palace, e main set was placed at one end of the studio, and the sound n this set was picked up by an omnidirectional moving-coil icrophone mounted on an extensible boom of the type already n use in the film industry. When an orchestra was needed, it as placed at the far end of the studio, and fixed microphones ere used for the orchestra or for small subsidiary sets in other arts of the studio. The small size of the studio made it mpossible to use more than one boom. A need for some form f directional microphone was soon felt, in order to pick up the oices of singers with the minimum of unwanted sound from he orchestra, and thereby facilitate the balancing of voice and rchestra at the sound control desk. A ribbon microphone with figure-of-eight polar diagram was in use in sound studios, but t could not be used on a boom since it produced rumbling noises hen moved, owing partly to the draught created and partly to echanical vibration. It thus became necessary, on musical rogrammes, to accept the fact that the vocalist's microphone as also the main orchestral microphone, and to balance as ell as possible with fixed microphones picking up particular roups of instruments.

Soon after the war, a type of cardioid microphone, which ould be mounted on a boom, came into use in the television tudios. It produced the cardioid polar diagram by combining he outputs of moving-coil and ribbon elements, and was ufficiently light, robust and free from wind noises for boom ounting.

When the Lime Grove studios came into service, the much reater size of these studios made it possible to use more than ne microphone boom, which facility was of great value in roductions with a number of sets.

When a microphone is mounted on a boom, or held in an terviewer's hand, or used in other ways where a breakdown ould entail an interruption of the programme, robustness and eliability are just as important as high quality. On the other and, where the outputs of several fixed microphones are mixed a dance band or orchestral broadcast, or where two fixed icrophones can be slung together in the main position for an rchestral transmission, the microphones can be chosen for uality above all other considerations. The high-quality ribbon microphone developed by B.B.C. engineers is frequently used in these circumstances. The microphones used in television generally fall into one of these two main classes.

In televising opera and other very-large-scale productions, the singers and orchestra are sometimes placed in different studios, and the outputs of the microphones in the two studios are mixed to form the transmitted sound signal. In order to enable the singers to hear the orchestra, the sound of the orchestra must be relayed by loudspeakers in the singers' studio. If this music is diffused indiscriminately throughout the studio, the final sound transmission must inevitably be coloured by a delayed and distorted version of the orchestral music picked up by the singers' microphones. A new arrangement of loudspeakers and microphones has recently been devised to avoid this coloration. Directional loudspeakers, facing towards the singers, are mounted on the trolleys of each of the microphone booms. The cardioid microphone at the end of the boom has its axis of maximum response facing the singers, so that its line of minimum pick-up is facing rearwards towards the loudspeaker. Since each loudspeaker is directional, it does not feed much sound to other microphones in the studio. In this way the singers' voices are picked up separately and can subsequently be 'balanced' with the sound from the orchestra.

The need has long been felt for a more sharply directional microphone which would single out such sounds as a tap dancer's feet or speech from an individual member of an audience. The parabolic microphone used at cricket matches and other outside broadcasts is too unwieldy for television studio use, but a new directional tube microphone, manufactured in Germany, has recently been tested operationally with some success. It consists of a tube, 1 m long, having a slot running the whole of its length, through which the sound enters. The slot is covered with fabric of suitably graduated acoustic resistance, and the operation of the microphone is analogous to that of a Beverage aerial.

Considerable use is made in the television service of the B.B.C.-designed radio microphone.[19]

(10.2) Acoustic Treatment of Studios

When the television service started, the accepted practice in sound broadcasting was to have different studios specializing in particular types of programme, and therefore the acoustics of each studio could be designed for one type of programme. This policy could not be adopted at Alexandra Palace because there were not enough studios, and an attempt had to be made to produce 'compromise' acoustics. Studio A had a reverberation time varying from 1 sec at 60 c/s to 0·5 sec at 8 kc/s. This proved too 'live' for the small dramatic productions of the time, in which the aim was to create an impression of intimacy in the dialogue of the few actors taking part. On the other hand, the studio was too 'dead' for the larger musical productions, especially when those involved crowding a large number of sound-absorbing human bodies into an area which was not really large enough for the purpose.

When the Lime Grove building, with its much larger studios, was acquired, the acoustic treatment of each studio was planned with a particular type of programme in mind. Studio D was made suitable for Drama; Studio E for Children's Hour, which is drama with a little music; Studio H is similar to Studio E; and Studio G is suitable for light entertainment in which there is a considerable amount of music of the popular kind.

(11) STUDIO AND CONTROL ROOM EQUIPMENT

The original control and preview facilities at Alexandra Palace were, by present-day standards, very restricted. The producer and vision mixer did not have a continuous display of

the pictures from all the working studio cameras. They had only a 'transmission' and a 'preview' monitor, and were completely dependent on the camera control operators for the supply of the correct picture signals at the appropriate moments in a production.

As each of the Lime Grove studios was brought into use, the vision control rooms were equipped with enough picture monitors to cover all the working studio cameras, and also to preview any telecine or outside-broadcast inserts which might be required.

Following tests which had been carried out at Alexandra Palace, radio equipment of B.B.C. design was provided at Lime Grove to enable the producer in the soundproof control room to give instructions and cues to the studio manager on the studio floor. This operated on frequencies in Band I, and the super-regenerative a.m. receiver complete with battery was small enough to be slipped into the studio manager's pocket. The headphone lead was used as a receiving aerial. Subsequently a transmitter, only slightly larger than the receiver, was designed to enable the studio manager to speak to the producer. The microphone lead was used as the transmitting aerial.

In 1954 the Lime Grove Central Control Room and its associated areas were completed. The vision and sound signals from all studio and telecine sources in Lime Grove were passed through this control room, and thence through a quality check room to the Television Switching Centre at Broadcasting House, which was then the hub of the B.B.C. television network. A small presentation studio was located alongside the Central Control Room, to enable announcements to be made between the programmes—at short notice if necessary.

'Genlocking' also became available at this time. This enables the waveform generators at Lime Grove to be synchronized with picture signals entering the building from an outside source. By this means, incoming pictures can be mixed with those produced in Lime Grove, and 'cutting' between the internal and external picture sources can be carried out without causing a momentary loss of synchronization in receivers.

By 1955, the gradual improvement in the performance of the permanent simultaneous broadcast links and their associated return circuits had reached the stage where it became possible to adopt continuity working, which had already been standard practice in sound broadcasting for some years. All vision signals were routed through the Central Control Room at Lime Grove—whatever the point in the United Kingdom at which they originated—and afterwards distributed to the national transmitter network. This meant that an outside broadcast from, say, a place near Kirk o'Shotts would be relayed a total distance of some 900 miles before being radiated by the Kirk o'Shotts transmitter. The procedure was nevertheless justified, because any alternative arrangements involved simultaneous mixing operations in different parts of the country. This would have been very cumbersome and would have become progressively more so as the number of transmitters and potential programme sources continued to increase.

(11.1) Production Lighting

Until the conversion of the Riverside studios for television was completed in 1956, no B.B.C. studio had really adequate lighting control facilities. The arrangements for positioning, switching and dimming the studio lighting were crude and could not fully provide the rapid setting-up of lighting, or the rapid lighting changes during the programmes, demanded by television. This applied particularly to Lime Grove, where the lighting facilities remained largely as they had been in the days when the studios were used for film production. In film-making, long time intervals between 'takes' are customary, and the setting-up and control of the lighting can be carried out at a more leisurel pace.

In both the Riverside studios the light sources are suspende from a large number of horizontal tubular battens. Eac batten normally carries four luminaires and is suspende from the ceiling by a motor-driven hoist. This enables th luminaires to be raised or lowered at will to any height betwee ceiling and floor level. All the hoists in each studio are co trolled from a single control desk on the studio floor.

The switching and dimming of the lights, either individuall or in any desired prearranged groups, is carried out from ligh ing consoles in the control room. In order to provide exper ence which would assist in planning the new Television Centr the two studios were provided with different systems of dimmin In No. 1 studio the dimming is by variable resistances, and few auto-transformers, both operated mechanically throug electromagnetic clutches. In No. 2 studio the light output of luminaires is controlled electronically by xenon-filled thyratron Experience with the two studios has shown that the mechanic system is preferable for television use.[6]

The preferred placing of the three studio control roo (vision, sound and apparatus) is to have all three adjacent to o another on the same floor level, with windows to give the visio and sound control rooms a direct view of the studio flo This was achieved for the first time at the two Riverside stud

(11.2) Special Effects

In 1951, back projection came into use as an alternative (und suitable conditions) to painted scenery. With this system th background in a studio scene is projected on to the back of translucent screen, taking care that the perspective of the pro jected scene matches that of the actors and 'properties' appeari in front of it. Moving backgrounds can also be projected, to gi effects such as a street scene with passing vehicles; in orde avoid strobing or flicker effects it is necessary to run the pro jector in synchronism with the television field frequency, and phase the projector driving motor so that the film pull-dow does not occur while the image of the back-projected scene being scanned by the camera.

Late in 1952, two similar and very useful techniques know as 'inlay' and 'overlay' were introduced into the studios.[20]*

The inlay process enables the technical operations staff make a 'hole' of any desired shape or size in the picture fro one camera channel, and to fit exactly into this 'hole' anoth picture from an entirely separate source. Thus, for examp an exterior view of a building may show an open window, an the area of the window can be filled in by a picture (comi from another camera) showing a person in a room interio This is done entirely by electronic and photo-electric means. flying-spot raster is produced on a vertically mounted tube wi its flat screen facing upwards, and light from it is intercept by a photocell. If any opaque piece of masking material is la on top of the raster, the light from each scanning line will interrupted as the flying spot passes under the masking materi and the photocell output will be interrupted at the same tim This interrupted photocell signal can be made to operate electronic switch which instantaneously transfers the visi output from one picture source to another. Thus the area to inlaid can be completely determined by the shape and positi of the masked area on the raster. The system will immediate follow any changes in the masking, and numerous variants of t basic idea of a movable mask are used to give different spec effects.

The overlay system uses the same electronic switching, but this case the switching signal is derived from an actual pictu

* British Patent Nos. 743402, 748822 and 489717.

hus if an actor in light clothing performs in front of a black ackground, the resulting picture signal will change from near-lack to near-white as the figure of the actor is scanned. So ng as the transition in voltage is well defined, this signal can perate the electronic switch and allow the picture of the actor be superimposed on a background from another picture ource, without making the actor appear transparent as he would a normal 'mix' were used. The movement of the switched ea will, of course, automatically follow the movement of the rtist's outline. Since the foreground and background pictures e taken by different cameras they can be of different relative zes. Striking effects can thus be produced; for instance, a ancer can be made to appear dancing on the rim of a wine lass or on the palm of a man's hand.

The overlay system is not always completely successful because ghting is restricted by the need to keep the background dark, nd unavoidable shadows on the subject produce spurious witching signals which cause unsteady tearing effects on the omposite picture. Experiments have therefore been carried out y the B.B.C. using a colour-separation method of obtaining the witching signals. This produced the desired effect most suc-essfully, but the system requires the use of three cameras and pecial lighting.

A new design of inlay equipment has been installed at the elevision Centre using a vidicon camera channel in place of the ying-spot technique currently in use for the generation of the witching signal.

(11.3) Video Transmission Chain

Continuous attention has been given to improving the per-ormance of amplifiers and other apparatus comprising the video ransmission chain. This development has reached such a point the Television Centre installation that the stability of the pparatus renders it unnecessary to put individual level adjust-nent on any amplifier in the chain from the studio camera output the main outgoing transmission terminal. This has been chieved by equalizing all internal tie lines accurately to a given oss and making the amplifiers with a very high order of gain onstancy.

The routing of video signals from the various sources, such as elecine machines, video-tape machines and outside broadcasts o and from the various studios is carried out with rotary witches. The operations of routing particular programmes ources to particular destinations are carried out by one or two perators from a central remote control position. This arrange-nent gives very great flexibility of control.

In the design of more complex video equipment such as syn-hronizing-signal stabilizing amplifiers care has been taken to void distortion being introduced in the television signal because f defects present in the incoming signal. For example, the pparatus does not increase the visibility of any noise that may e present on the incoming signal.

All apparatus at the Television Centre has been designed to e capable of handling a colour-television signal in order to rovide for future developments. All apparatus in the main ransmission chain uses electronic valves, but development has ow reached the point where it is possible to introduce some video pparatus using transistors.

(11.4) Film Studios

Early in 1956 the Ealing Film Studios were purchased by the B.B.C. and have since been used for the making of documentary lms, film inserts into studio programmes, and other films apart rom those needed by the News and Newsreel Service. The film eviewing, cutting, dubbing and sound-recording facilities have been substantially modified and extended since the B.B.C. took over the Ealing premises.[21]

No separate film studios have been provided outside London, but equipment for the shooting and editing of film and to a limited extent its processing has been provided in the main regional centres—as well as equipment for the transmission of films.

The B.B.C. is the biggest single user of film in the world, and the film department produces, in terms of footage, the equivalent of some 140 full-length feature films each year.

(11.5) Film Dubbing

In film making for television the word 'dubbing' is used for the process of transferring the sound accompaniment of a film from one of the many sound recording media to another, or for adding a commentary to silent parts of the film.

Recording or reproducing machines for all the forms of sound recording involved in the dubbing process must be provided. There are at present some 14 variations which may be encoun-tered, embracing both 35 and 16 mm film. It is also necessary to provide the sound accompaniment of films for export (trans-criptions) in a variety of different forms.

Originally, the recording and reproducing facilities were dis-persed among various B.B.C. premises, but a comprehensive sound transfer suite was built in the television film studios at Ealing and brought into use in 1958. This is a room containing equipment for recording and reproducing any of the forms of optical or magnetic sound tracks likely to be encountered, together with disc reproducers and arrangements for running any combination of the machines in synchronism with one another. The film dubbing theatre includes facilities for observing a pro-jected picture and recording a spoken commentary or other sound accompaniment, and for this the commentator is pro-vided with a microphone so placed that he can view the film and time his remarks to coincide with the action. At the same time, the microphone must be positioned so that it does not pick up unwanted sounds in the dubbing theatre. The control desk in the dubbing theatre enables sound signals from film, tape, disc or microphone to be faded up at the appropriate moments, and cueing facilities are available between the producer and the commentator and operators.

(12) DEVELOPMENT OF TELEVISION POINT-TO-POINT LINKS

When the B.B.C. television service started, the transmitters and the studios were all in the Alexandra Palace building, and it was possible to transmit programmes without using any external vision links.

As soon as the television service had restarted after the war, plans were made in co-operation with the Post Office for extend-ing the service to the provinces in accordance with the recom-mendations of the Hankey Committee. The first regional transmitter was planned to be at Sutton Coldfield, near Bir-mingham, and both radio and cable links were installed by the Post Office between London and Birmingham, in order to provide experience of the relative merits of the two types of link.[22, 23]

The design of the London–Birmingham cable link was in-fluenced by the Report of the Hankey Committee, in which it was recommended that development should be planned on the assumption that a system of television employing perhaps 1 000 lines would be introduced in due course. This would presuppose a bandwidth of the order of 12 Mc/s. The possible introduction of colour television at a later date was also considered, and since no method of bandwidth-saving had then been developed, pro-vision was made for the cable to transmit a bandwidth of the order of 30 Mc/s without difficulty if the need arose.

The next stage in the development of the simultaneous broadcasting system was the planning of a coaxial cable link from Birmingham to Manchester and the Holme Moss high-power transmitter. By this time it had been agreed that the 405-line system should remain in use indefinitely, and the provision for video bandwidths greater than 3 Mc/s was dropped. A 2-way link, using standard 0·375 in coaxial tubes, was provided. For the longest run of the link—from Birmingham to Manchester—the two television coaxial tubes shared a lead-covered cable with four other 0·375 in coaxial tubes and a number of telephone pairs. The loss in the coaxial tubes was 6·5 dB per mile at 3 Mc/s, compared with 2·5 dB per mile at the same frequency for the 0·975 in-diameter London–Birmingham cable. On the 84-mile run from Birmingham to Manchester 16 repeater points were provided. As with other coaxial systems, it was necessary to translate the frequency band upwards in order to avoid having components in the very-low-frequency range. At the same time, however, the greater attenuation made it impracticable to transmit components as high as the 7 Mc/s reached on the London–Birmingham link, with the number of repeater points used. In this case the video band was transmitted with a carrier at 1·056 Mc/s, an upper sideband cut off at 4·056 Mc/s and a vestigial lower sideband extending to 0·5 Mc/s below the carrier.

Some quadrature distortion is inevitably produced in any asymmetric sideband system, but in this case it was minimized by the use of negative modulation, which meant that the picture portion of the signal did not extend above 70% modulation. The presence of the carrier and most of the signal power at the lower end of the transmitted band meant that high linearity was needed if patterns caused by harmonics were to be avoided, but this was aided by a new type of modulator valve, the CV 2127.

A similar cable system was used for the link from London to the Wenvoe high-power transmitter in South Wales.

The planning of the Post Office links for the extension of the distribution system to the Scottish high-power transmitter at Kirk o'Shotts was started in 1950, and it was decided that s.h.f. radio links, with frequencies of about 4 Gc/s should be used. After some uncertainty whether seven or eight repeater stations would be necessary, preliminary tests showed that seven would give adequate reliability. A route east of the Pennines was chosen, in order to serve the projected medium-power transmitter at Pontop Pike, near Newcastle upon Tyne. The basis of planning was a maximum line-of-sight path length of about 40 miles, with 'first Fresnel zone' clearance of intervening obstacles. After the preliminary tests, however, the seven repeater points actually chosen involved one 47-mile path with a 200 ft aerial tower at one end.

The B.B.C. at present rents from the Post Office some 2 100 channel-miles of vision circuits, of which 900 miles consist of fully engineered 2-way links (i.e. 1 800 channel-miles) and 300 miles are 1-way links.

Some of the stations in the network rebroadcast signals which have been received directly from another television broadcast transmitter. The use of a tall mast on a high site at the rebroadcasting transmitter usually makes it possible to mount a receiving aerial giving a better signal pick-up than any local viewer could obtain. Alternative receiving aerials at different heights enable a choice to be made when receiving conditions are difficult. In some instances, where a satisfactory direct pick-up cannot be obtained at the site of the rebroadcasting transmitter, the broadcast is received at some suitable intermediate point and conveyed the remainder of the distance to the rebroadcasting transmitter by a short s.h.f. or v.h.f. link. Links of all these types equipped and maintained by the B.B.C. total some 950 miles.

(12.1) Testing of Television Links

Considerable progress has been made in the development improved methods of testing vision circuits for routine ma tenance. The major distortions which need to be assessed determining the standard of quality of a television transmissi link or network are those of attenuation and phase with resp to frequency, and also linearity and noise.

Although accurate steady-state measurements of attenu tion/frequency and group delay/frequency characteristics can made—considerable effort was expended by the B.B.C. designing equipment for this work just after the war—th measurements are not suited to routine maintenance work, a the emphasis was originally placed on waveform testing, usi rectangular pulses of different durations to reveal on an oscil scope the characteristics of the circuit under test. This syst suffers from certain limitations in practice:

(i) In order to reveal distortions occurring towards the upper of the frequency band, the test pulse must have a rise time sm compared with the period of the upper-frequency limit of the li The frequency spectrum of the test pulse consequently extends above the highest frequency of interest, and undue emphasis m be given on the received waveform display to distortions which of no importance to the transmission of television signals.

(ii) The waveshape, being dependent on valve circuits, can cha with time, and different waveform generators may produce differ waveshapes.

(iii) The television picture signal has a limited spectrum, i.e. i of a different nature from the test waveform.

(iv) Accurate measurement of short rise times is not easy.

In order to overcome these disadvantages, and for ot reasons concerning the relation between waveform distort and subjective distortion of the picture, and the prediction of performance of tandem connections of links, a pulse of nomina sine-squared shape was developed, in particular by the P Office.

This test waveform, combined with a suitable rectangu pulse and known as a 'pulse and bar' test signal, has a limi spectrum and enables measurements to be made without int ducing any irrelevant information, while giving a more string test of upper-frequency amplitude and phase distortion than previous unlimited-spectrum test waveforms; in addition, waveform is reproducible with certainty by different generat and measurements are in general easy. It is now in general and will soon supplant entirely the older method.

Linearity has been for many years estimated by inspection linear sawtooth waveform, or by measuring the received amp tudes of a synchronizing pulse waveform with varying amou of a full-line signal in the picture period. Considerable discuss on the measurement of linearity has taken place on an in national level without full agreement being reached, but B.B.C. is now bringing into use the internationally recogni staircase waveform with five steps from black to white level.

Noise has for many years been measured as the quasi-peak peak noise amplitude displayed on an oscilloscope. T method gives very inconsistent results owing to such varia factors as trace brightness and ambient illumination, and it been felt for some time that a meter display method would much superior. Such methods are now being used for measur random continuous noise in the presence of the television sign and are giving much better reliability. It is, of course, necessary to use oscilloscope methods for measuring rand impulsive noise.

(13) OUTSIDE BROADCASTS

For the first few months of the television service, the range which cameras could operate was limited to points in the A

dra Palace grounds which could be reached by a 1 000 ft run camera cable from the control room. This was the maximum gth of cable for which it was then considered practicable to ovide a compensating delay for the timing pulses.

The need to venture farther afield was realized from the start, wever, and two mobile outside-broadcast units were brought o service before the war. The first outside broadcast was of Coronation procession of King George VI on 12th May, 37.

Each of the outside-broadcast units comprised four vehicles, out the size of large motor coaches. One vehicle contained equipment for a complete mobile control room (known as m.c.r.), including working and standby waveform generators, ntrol equipment for three Emitron or Super-Emitron cameras, eview and transmission picture monitors and mixing equipnt for both vision and sound.

The other vehicles in the outside-broadcast unit contained, pectively, a vision radio-link transmitter, having a peak white wer of 1 kW and operating on a frequency of about 60 Mc/s, extending ladder for raising the transmitting aerial to a height some 80 ft and a Diesel-generator set for use where no public wer supply was available.

On the occasion of the Coronation, and for other outside oadcasts from the Central London area, the vision signals re conveyed to Alexandra Palace by a vision cable circuit stalled by the Post Office. This was a balanced-pair cable, as was intended to pass video-frequency signals directly through without using a modulated carrier to translate the signals to a gher frequency band. The balanced-pair system enabled erference at very low frequencies to be cancelled out. At out the same time a technique was developed by B.B.C. gineers for equalizing telephone cables to enable vision signals be passed through them over limited distances, in order to k up with the nearest available point on the balanced-pair twork. This technique is still widely used. The loss at 3 Mc/s fore equalization, when using this system, must not normally ceed 66 dB, and this determines the maximum possible disnce, which varies—according to the type of telephone cable in e—from three-quarters of a mile to two miles.

These outside-broadcast units were widely used throughout e London area before the war and for a time afterwards. A dio-link receiving point was established on Highgate Hill to ercome interference, which was liable to arise when attemptg to receive direct at Alexandra Palace in the presence of the rong signals from the local transmitters. From Highgate to lexandra Palace the vision signals were carried by a section the balanced-pair cable mentioned above.

(13.1) Post-War Outside-Broadcast Equipment

All the later outside-broadcast units incorporate the same asic facilities as these early vehicles. The newer units are more mpact, the layout of the m.c.r. has been made much more nvenient from an operational point of view, and the radio nks use higher frequencies. (The development of radio links r outside broadcasts is described in Section 13.6.)

The first m.c.r. to be brought into service after the war used x-wheeled articulated vehicles. It was thought that the most ficient way of working would be to uncouple the trailer portion the vehicle after it had been driven to an outside-broadcast te, and thereby release the mechanical horse for use elsewhere. was found, however, that the awkwardness of these vehicles, pecially on the more difficult sites, outweighed the expected lvantages, and all the more recent m.c.r.s have been built into lf-propelled four-wheeled vehicles.

When broadcasts are made from sites where it is not possible park the van at a suitable location, the complete equipment can be removed from the vehicle and set up in a nearby building or even in a tent. When this is done, each unit of the equipment is disconnected from the cables in the van and a duplicate cable harness is used to link the units wherever they have been set up.

Each outside-broadcast unit includes a number of auxiliary self-propelled vehicles equipped with radio links and aerials, and a vehicle in which the cameras, camera cables and ancillary equipment such as picture monitors are transported to the site. Mobile power-generating equipment is available if required.

(13.2) Use of Image-Orthicon Cameras

The first mobile control rooms to be brought into service after the war used c.p.s. or image-iconoscope cameras. However, as soon as image-orthicon cameras, using 3 in tubes, came into use, it became clear that the much greater sensitivity of this type made it the best camera for outside broadcasts, and it has since become the only type of camera used for this purpose, except in special applications where the need for extreme compactness or lightness makes it essential to use vidicon cameras. Several of the 3-camera outside-broadcast units now in service use the new cameras with 4½ in image-orthicon tubes; it is felt that the advantages of higher signal/noise ratio and better tonal gradation more than outweigh the disadvantages of greater bulk and weight compared with the earlier 3 in tube cameras.

(13.3) Special Applications

A television broadcast from a submarine at sea—which was first carried out on 16th June, 1956—is an example of the type of transmission in which vidicon cameras have been used because of their compactness and lightness. In the type of submarine from which these broadcasts have hitherto taken place, there is room for an image-orthicon camera in the forward torpedo compartment, but the production requirements called for another camera fixed to the periscope, and a third which could be moved to different parts of the ship. For these two positions vidicon cameras were essential, in spite of the difficulty of obtaining enough lighting to enable them to work satisfactorily.

(13.4) Radio Camera

Another application demanding the use of a small, lightweight camera is the so-called 'radio camera'. The first radio camera to be used by the B.B.C., which was purchased from a French manufacturer in 1957, was first used on a live transmission[9] in January, 1958. It consists basically of a vidicon television camera and control unit, a transistorized waveform generator, a radio transmitter, and battery power supplies, with a total size and weight which can be reasonably carried by a single person. With a 100 mW output transmitter, giving an average range of 300–400 yd, the total weight is just under 31 lb. With an additional unit giving a transmitter power output of 5 W and a range of two to three miles, the total weight (including additional batteries) is about 49 lb. Given a lighting level within the capabilities of a vidicon tube, this equipment opens up possibilities of transmitting live pictures under conditions where this would otherwise be impossible.

(13.5) Roving Eye

A new type of mobile unit, which is referred to as the 'roving eye', was used on a programme for the first time on 17th January, 1954. This was a single image-orthicon camera channel mounted in a van-type vehicle of medium size, together with a petrol-electric generator, transmitters and aerials for sending picture, sound and control signals to a fixed receiving point, and also a commentator's microphone. The vision radio link used a frequency in Band III, and a gyro compass was used to

keep the transmitting aerial aligned on the receiving point. With an installation of this type, live pictures and a commentary could be broadcast from the vehicle while it was moving—hence the name 'roving eye'. Operational experience with this unit soon showed, however, that its roving capabilities were of secondary importance compared with the ability to provide, at very short notice, either an additional fixed camera—with its own radio-link equipment—at outside broadcasts where a 3-camera mobile control room was in use or, alternatively, an independent single camera unit. At race-meeting broadcasts, for example, there is often a need for a camera near the starting gate, too far away from the 3-camera mobile control room to be reached by normal 1 000 ft camera cables. Recently, however, some mobile control rooms have been provided with sufficient delay compensation to permit the use of 2 000 ft camera cables, and this more often enables the remote camera position to be operated with a camera from the mobile control room.

In December, 1956, a second roving-eye vehicle, equipped with two working cameras, came into service.[24] Careful planning of the layout, together with the use of a 'forward-drive' chassis, made it possible to carry the extra camera together with all the facilities available in Roving Eye I, in a vehicle 18 in shorter than its predecessor. An f.m. vision radio link, working at about 660 Mc/s, was used in this case. The provision of two cameras, together with vision mixing equipment, has greatly increased the usefulness of this vehicle. In addition to being used in conjunction with a 3-camera mobile control room it is often used alone as a stationary 2-camera outside-broadcast unit, with one or both cameras working away from the vehicle on extension cables.

(13.6) Radio Links

Where suitable Post Office cable circuits do not exist, B.B.C. point-to-point radio links are employed, using mobile equipment to send the vision signals from outside-broadcast sites to the nearest convenient injection point on the permanent network.[25] This equipment is based in London and at six regional centres and enables vision signals to be relayed from almost any point in the British Isles.

The history of these links is, for the most part, one of progressive increases in frequency. The original pre-war links working on a frequency of about 60 Mc/s became difficult to use as additional channels in Band I became occupied by B.B.C. television stations, and eventually had to be abandoned. This was a pity because their ability to operate where an optical path was not possible was unrivalled.

The next step was to use equipment operating on about 200 Mc/s. A 14-ton vehicle was equipped with the transmitter, a Diesel-generator and a fire escape ladder extending to 120 ft for elevating the aerial. These links also gave good service. The increased liability to multi-path signals on these frequencies was largely compensated by the more highly-directional aerial arrays of reasonable dimensions which it became possible to use. However, these frequencies, in turn, had largely to be abandoned when Band III began to be used for television broadcasting, the purpose for which the band was originally allocated by international agreement.

Several years before Band III television broadcasting started in 1955 the use of s.h.f. links had become well established for outside broadcasts. During 1948–50, radio links operating in the frequency range 4·4–4·8 Gc/s became available commercially and permission was obtained from the Post Office to use these frequencies on a temporary basis. Additional demands by other services have now made it necessary to move to the region of 7·05–7·3 Gc/s. (In the present state of the art, frequencies above 10 Gc/s are difficult to use on account of increased attenuation caused by fog, rain or snow.) As many as five s.h.f. links have been used in tandem to relay picture signals fr[om] exceptionally difficult sites.

It is not always possible to find a suitable site close to [the] mobile control room for the first microwave link, and so[me] form of short-range link between the microwave link and [the] m.c.r. must be provided. For this purpose the new u.h.f. f[...] links developed by the B.B.C. are often used.[26] These work [in] Band V (522–660 Mc/s), and the lower frequency, as compa[red] with super high frequency, offers important advantages in [this] particular application.

(14) DOMESTIC TELEVISION RECEIVERS

(14.1) Pre-War Receivers

Before the war, with only a single transmitting station [in] operation in Great Britain, a fixed-tuned receiver was all t[hat] was necessary. Since it was thought that the maximum ra[nge] of the Alexandra Palace station would be no more than ab[out] 30 miles, representing a field strength of $\frac{1}{2}$–1 mV/m, receiv[ers] were designed with only moderate sensitivity. In many mod[els] adequate gain was achieved by the use of t.r.f. circuits; so[me] manufacturers, however, preferred to use a superheterod[yne] circuit with an intermediate frequency in the region [of] 10–13 Mc/s, since the necessary gain was easier to achieve [at] these frequencies. The majority of receivers used picture tu[bes] of 9 in diameter or less, and, with an e.h.t. of the order [of] 4–5 kV, the picture brightness was barely adequate for view[ing] in daylight or in normal room lighting. The e.h.t. supply [was] normally obtained from a valve rectifier fed from the ma[ins] transformer. Usually the same transformer provided [the] receiver h.t. supply and a low-voltage supply for the va[lve] heaters, which were fed in parallel. The time-base circ[uits] frequently made use of gas-filled valves, although there wa[s a] wide variety of hard-valve time-bases. Both electrostatic [and] electromagnetic scanning circuits were used, although by 1[939] electromagnetic deflection and focusing had become the m[ost] common.

(14.2) Post-War Receivers

When the television service was restarted after the war, [no] major difference in receiver design was at first appare[nt]. Improvements in valves and circuits did, however, result in [an] increase in sensitivity, which was important because the ra[nge] of the Alexandra Palace station had been found to be co[n]siderably greater than the 30-mile radius originally envisag[ed]. In the absence of severe local interference, reliable recepti[on] was being obtained at a field strength of 100 μV/m or l[ess], corresponding to a range of at least 50 miles.

A number of basic changes in receiver design subsequen[tly] took place.

(14.3) Tuning Arrangements

With the opening of the second B.B.C. television transmitt[ing] station at Sutton Coldfield in 1949, it became necessary to p[ro]vide receivers which could be tuned to Channel 4. Previou[sly] only Channel 1 had been used.

This problem was tackled in a number of ways, but the m[ost] elegant solution was to provide an interchangeable sub-cha[ssis] in the r.f. section of the receiver. The opening of further B.B[.C.] stations using additional channels in Band I led to the gene[ral] adoption of some form of variable tuning which enabled receiv[ers] to be tuned to any of the five channels. Superheterod[yne] receivers were then universally adopted.

A further basic change took place in 1955 with the bring[ing] into use by the I.T.A. of frequencies in Band III for televis[ion]

padcasting in the United Kingdom. The switched turret ner made its appearance, and, according to the design, pro- ed for reception on any number of channels up to 13. Some nufacturers preferred to use switched incremental inductance ning providing a similar number of channels.

The B.B.C. v.h.f. sound broadcasting service was also intro- ced in 1955, and within a year or so some television receivers luded provision for receiving this service on frequencies in nd II; one or more of the switched tuning positions were ear- rked for this purpose. In addition to the switched tuning, it s usually been found necessary to provide a fine tuning control operation by the user; some models have recently appeared in ich improved frequency stability has made this unnecessary cept as an internal preset adjustment. The low intermediate quencies (9–16 Mc/s) used in pre-war receivers, and for a nsiderable time in post-war receivers, have been discarded. A ndard intermediate frequency band of about 34–38 Mc/s has w been adopted by almost all manufacturers.

(14.4) Cathode-Ray Tubes

One feature of post-war receivers has been an increase in the e of the picture, which has usually been achieved by increasing e diameter of the cathode-ray tube. This led to an undesirable crease in the size of the receiver, and various means were opted to overcome this. It became possible to reduce the dth of the receiver when cathode-ray tubes with rectangular ce-plates were introduced, making almost the whole of the ontal area of the tube available for the picture. The depth the receiver from front to rear, always something of a problem, as reduced by designing cathode-ray tubes with wider deflection gles so that the length of the tube could be drastically reduced. eflection angles have progressively increased, first to 70° then 90°, and in the past year or so to 110°. The front-to-back mension of the modern 21 in tube is now less than that of a n pre-war tube. The introduction of tubes with 110° deflection gles has enabled the depth of the complete receiver to be duced markedly, giving a 'slim' effect.

Tubes up to 21 in diagonal are commonly used for domestic ewing, but more than 80% of the receivers now sold have in tubes.

In spite of the considerable increase in picture size, improved chniques have enabled the picture brightness to be increased so. This has involved an increase in e.h.t., which is now mmonly of the order of 15–20 kV. The use of aluminized cture tubes is almost universal, although recent developments tube design have made the difference in performance between uminized and non-aluminized tubes less than it once was.

Improved viewing in daylight or under normal room-lighting s been aided by the use of neutral density filters incorporated ther in the safety-glass screen mounted in front of the tube or the face-plate of the tube itself.

In 1949 a projection viewing tube and a Schmidt optical unit r use with it became available commercially at a competitive ice. The screen diameter of the tube was about 2½ in, and a ecially-deposited gelatine film was used for the aspherical rrector lens in the Schmidt optical system. An e.h.t. of kV was required, and a special e.h.t. supply unit was pro- ced, in which a positive surge of over 8 kV was developed the anode of a pentode by feeding a pulse to its control grid a repetition frequency of about 1 kc/s. This surge voltage was en fed to a tripler circuit in which the network of high-voltage odes and capacitors was enclosed in a sealed oil-filled container.

This projection system was incorporated in a number of mestic receivers, in which the picture was projected on the ck of a translucent screen at the front of the receiver. ceivers of this type have not, however, come into very wide- spread domestic use. A picture size of 20 in × 15 in or 24 in × 18 in was readily obtained.

(14.5) Power Supplies

In most receivers the bulky and costly mains transformer quickly disappeared after the war. This was made possible, first, by the development of e.h.t. supplies obtained by rectification of the high voltage pulse produced in the line-scanning circuit during the flyback period, and secondly, by the introduction of the a.c./d.c. technique in which the valve and tube heaters are connected in series instead of in parallel. This technique was made possible by the development of valves which would operate efficiently with an h.t. of 200 V or less, resulting from direct rectification of the mains supply.

(14.6) Scanning Circuits and Deflection Systems

Gas-filled valves were formerly used in the time-base circuits, but these have been replaced by hard valves. The need to generate considerably more power in the scanning circuits used with larger screens and wide-angle tubes has led to a variety of designs incorporating means of reducing distortion and increasing linearity. The major problems have arisen in the line-time-base circuits, for which it has been necessary to design special valves of high efficiency and reliability for the output stage. The use of a so-called 'efficiency diode' is almost universal to provide the necessary damping of the line-scanning circuits so as to prevent 'ringing', caused by oscillations of the scanning waveform during the flyback period. The energy transferred to the diode is recovered in the form of a rectified voltage which is used to boost the h.t. supply to the line output valve.

The use of interlaced scanning creates a number of problems in the design of time-base circuits and in the circuits used for extracting the synchronizing signals. There has been a steady improvement in the performance of receivers in this respect, and many of them now achieve a very high standard of interlace.

Much attention has been given to the design of deflection systems where new problems have constantly arisen with the increase in screen size and deflection angle. The scanning coil assembly is now commonly built on a ferrite core which has enabled the copper and iron losses to be substantially reduced.

For some years the use of magnetic deflection and focusing was universal, but improved designs of cathode-ray tubes for electrostatic scanning and focusing have recently made their appearance and have been incorporated in a number of receivers. The modern electrostatic tubes maintain focus at the edges of the picture better than magnetic focusing can do; there is also less defocusing on peak white.

(14.7) Receivers for Fringe Area Reception

A major problem arises in areas of low signal strength (owing to local screening and other factors) and areas in which varying propagation conditions cause severe fading. While these difficulties cannot be entirely overcome, there is much that can be done at the receiving end to provide reception of an acceptable standard. So-called fringe-area receivers have been produced for use in such circumstances. These have increased gain and usually incorporate flywheel synchronization circuits which help to keep the receiver in step with the transmitted pulses during periods when there is severe fading of the signal or during bursts of strong local interference. Line-tearing which produces ragged edges of the picture can, to some extent, be overcome in this way.

It is obviously important that receivers for use in such conditions should have an efficient a.g.c. system; there are two systems in general use. The first derives a control voltage from

the average signal level and is to be deprecated because it distorts tonal values of the picture. The second system generally referred to as 'gated automatic gain control' derives its control voltage from the back porch of the synchronizing signal, i.e. the period of black level (or suppression level) which follows the synchronizing pulse. The automatic gain control is then independent of the picture content. Some of these receivers also have a control which reduces the vision bandwidth under poor reception conditions.

(14.8) Limiters

Most modern receivers are provided with interference limiters in the vision and sound circuits to mitigate the effects of serious impulsive interference. Although these necessarily cause some deterioration in the received signal they tend to produce more pleasing results in the presence of severe interference. The sound limiter is normally fixed in the design of the receiver, but an adjustment of the vision limiter is generally provided for the user so that the best compromise can be made to suit conditions at the time.

(14.9) D.C. Component

In order that the tonal values of the picture shall approximate to those transmitted whatever the content of the picture, it is necessary to retain the d.c. component of the signal in the receiver. In areas where signal strength varies because of fading or other causes it is also necessary to use a well-designed, gated a.g.c. system deriving its control voltage from the black level of the signal, and not from the average value. Receiver design is inevitably a compromise between performance and cost, and there are few, if any, receivers which retain the full d.c. component. If all, or the majority of, receivers were designed to have a standard performance in this respect, the transmitting authority would at least know how its output was likely to appear on most domestic receivers. There would also be less restriction on the type of picture transmitted, giving the producers of the programme greater scope, if the d.c. component was retained in full.

(14.10) Aerials

In the early days it was customary for nearly every viewer to have at least a dipole aerial mounted outside the house, usually on a chimney. As the sensitivity of receivers increased, the practice tended to fall into disuse, and various makeshift indoor aerials were frequently employed in circumstances in which they were quite unsuitable, to the detriment of picture quality. The introduction of I.T.A. transmissions in Band III helped to make the viewer more aerial conscious because satisfactory reception without an adequate aerial was more difficult on these frequencies.

A considerable number of efficient aerial designs have been produced, including some that provide for reception of v.h.f. sound broadcasting in Band II as well as television in Bands I and III. The wide difference in appearance of the various structures erected above rooftop level is unsightly, and it is to be hoped that some more uniform and less obtrusive design will emerge.

(15) COLOUR TELEVISION

(15.1) Situation Prior to 1949

For many years it was assumed that 3-colour television would require the equivalent of three separate picture signals to be transmitted, each representing one of the colour components, and each having a definition equal to the accepted standard for a black-and-white picture. There appeared to be only two possible ways of doing this. One method would be a sequential system, in which the line and field scanning frequencies were trebled in order to allow three successive fields, each represent one colour component, to be transmitted in the time norm taken by a single field of the corresponding black-and-wl signal. The other method envisaged was a simultane system which would have the same scanning frequencies as equivalent black-and-white system, but in which three separ channels—each complete from camera to viewing tube—wo be provided. The three pictures received in this way would added to form the colour picture. Both systems had two fun mental disadvantages in that they required a total vision ba width three times that of a black-and-white picture with same definition, and, what might be equally important, neit system transmitted a signal which would enable existing receiv to produce a black-and-white picture of undiminished qua without any alteration of the receiver.

(15.2) Development of the N.T.S.C. System

On these assumptions, the advent of a public service of colo television broadcasting seemed as remote as did high-definit black-and-white television in the days before the use of the v. bands was out of the experimental stage. However, from ab 1949 onwards, various methods were proposed, and experime carried out, to overcome these fundamental difficulties. M of these developments took place in the United States, wh was the only country that could then afford any large-sc development of colour television, and the system develo by the American National Television System Committee approved by the Federal Communication Commission December, 1953. This system did not require any change in frequencies or bandwidths of the channels already in use black-and-white television, and it was compatible; i.e. the tra missions were in a form which enabled existing receivers, with any alteration, to produce a black-and-white picture with ne gible deterioration in quality. All this was achieved by tailo the system to the fundamental characteristics of human col vision in a highly ingenious fashion, and by an equally ingeni use of communication techniques. The N.T.S.C. system been fully described in the literature.

(15.3) Experimental B.B.C. Transmissions

In the United Kingdom, the B.B.C. developed an adapta of the N.T.S.C. system to the British 405-line system.[41] first transmission was made in October, 1954, and regular exp mental transmissions in colour have been made every autu and winter since 1955–56. A colour-television studio equipped at Alexandra Palace and experimental 'live' grammes, films and slides were transmitted. These tr missions were confined to the London area using, at first, Alexandra Palace transmitters outside normal programme ho and subsequently the Crystal Palace transmitters after station was brought into service. The purposes of these tr missions were as follows:

To provide a source of high-grade colour picture signals so help the industry to continue development work on colour recei

To enable further experience to be gained in the operation colour studio and other colour-television equipment.

To obtain further knowledge of the compatibility of the mod N.T.S.C. system.

To investigate further the problem of receiving and transmi colour pictures.

The data obtained from these experiments, in which the r industry co-operated, have been made available to the Televi Advisory Committee and have been published by the B.B.C. The main experiments came to an end in April, 1958, but at beginning of 1961 the B.B.C. was still continuing the tr mission of films and slides in colour to assist the industry in development of receivers.

The prolonged series of experiments has proved that the lapted N.T.S.C. system is fundamentally sound, fully comıtible, and capable of giving excellent pictures in colour.

The decision whether a public service of colour television ıould be introduced in the United Kingdom, and what form should take, rests with the Postmaster General, who has 'ailable to him the advice of the Television Advisory Comittee. The Committee's conclusions were published in its port[40] of June, 1960. On 9th December, 1960, the B.B.C. put ırward a proposal to start a limited experimental service in ılour in November, 1961, to coincide with the twenty-fifth nniversary of the start of B.B.C. television. The proposal was jected by the Postmaster General on 14th December, 1960; in ving his decision the Postmaster General said that the introıction of colour television must await the report of the Comittee on Broadcasting which had been set up in September,)60, to consider the future of the broadcasting services in the nited Kingdom.

(16) INTERNATIONAL EXCHANGE OF PROGRAMMES

In sound broadcasting, the exchange of programmes between ifferent countries was commonplace even before the period overed by this review, although it has expanded very conderably since then. The idea of extending these live proramme exchanges to the field of television was obviously ttractive as other countries began to establish television services.

Because of the wider bandwidth and the importance of ccurate compensation of both attenuation and phase disortion, the technical difficulties are much greater for vision than or sound, but for distances up to 1 500 miles or more they have lready been overcome. Indeed, the sound and control lines ssociated with the vision circuits often present more operational omplications because the commentary in sound may have to e carried in several languages simultaneously, whereas the icture is usually common to all the receiving countries.

The first major step towards international television relays vas taken in 1950 when the B.B.C. sent an outside-broadcast unit o Calais. Live pictures of a fête in the French town were rought to the screens of viewers in Great Britain by a series of 3.B.C. radio links of the type used for outside broadcasts. This vas a considerable technical achievement involving the use of our links in tandem.

(16.1) Standards Conversion

The B.B.C. and R.T.F. (the French Broadcasting Service) then onceived the idea of exchanging programmes between London nd Paris, but before this could be accomplished another major roblem had to be overcome. This was the difference in tandards of the French and British television services, which nade necessary some form of standards convertor.[44]

Various workers had written theoretical papers on this ubject suggesting that conversion might be achieved by so imple and obvious a proceeding as televising a picture on the riginal standard with a camera working on the new standard. Experiments on these lines—using the picture tubes and camera ubes in general use at the time—had given very unpromising esults. Early in 1952, however, the B.B.C. Research Departnent showed that this basic method was perfectly feasible if ertain precautions were taken.[45]

Experiments on such a convertor have revealed that three nain problems required solution before satisfactory results could e obtained. The first of these arises because the primary icture as displayed on the cathode-ray screen may be regarded s an intensity-modulated light spot rather than a continuous image. If the camera used in the apparatus can behave as a simple photo-electric cell, a signal will appear at the convertor output which corresponds to the brightness variations of the cathode-ray-tube spot. In this manner an unconverted component of the input signal will appear at the convertor output.

A second difficulty encountered when a camera is arranged to view a television image arises from the interference or strobing patterns produced when the scanning beam of the camera tube explores the line structure of the image to be converted.

The third fundamental problem is associated with any difference of field frequency which may exist between the two standards. Such a frequency difference must result in a cyclic variation of the vertical distance on the target of the camera tube separating the image of the cathode-ray-tube spot and the camera scanning beam. The properties of both the cathode-ray-tube phosphor and the camera tube are very important in this respect.

In the B.B.C. convertor the difficulties were overcome by the combined effect of (*a*) using a picture tube with an after-glow of about one-third of a field period, and (*b*) using a camera in which the output signal does not carry a superimposed signal representing the total amount of light, i.e. the image orthicon. The inter-line spaces in the picture to be converted must be filled in by line broadening or spot-wobble, in order to avoid a beat pattern caused by the difference in the two line scanning frequencies.

Using experimental convertors of this type, 819-line pictures from Paris were converted to 405-lines and broadcast by the B.B.C. in July, 1952. A similar type of convertor developed by French engineers was used to feed the 441-line service in Paris, which was then running side by side with the 819-line service. When the original pictures were satisfactory and the links were working properly, the method was quite successful.

(16.2) Eurovision

In June, 1953, pictures of the Coronation were relayed by France, the Netherlands and Western Germany. The main development of Eurovision started in 1954, by which time television services had begun in Denmark, Belgium, Switzerland and Italy. In the June of that year, the television services of these four countries, together with those of Great Britain, France, The Netherlands and Western Germany, were connected for a period of a month, during which each of the various services contributed, in turn, programmes for the benefit of all the other seven. Shortly after this, the European Broadcasting Union started to undertake the co-ordination of programme exchanges in both technical and programme spheres, and co-operation between the broadcasting authorities and the national P.T.T. administrations, already effective, was further strengthened. An International Technical Co-ordination Centre was established by the E.B.U. in Brussels.

In November, 1957, the B.B.C. began using $4\frac{1}{2}$in image-orthicon cameras (with their very high signal/noise ratio) in their convertors, and at the same time restricted the contrast range of the unconverted picture to about 5 : 1, in order to keep the voltage/light transfer characteristic almost linear. The net result has been a conversion of such high quality that it adds very little distortion to the final picture. Besides converting incoming 625- or 819-line pictures to 405 lines, the B.B.C. is now undertaking the conversion of outgoing signals from 405 to 625 lines. The horizontal definition of the 625-line picture need not be appreciably degraded by its being derived from a picture of a lower standard (as might be thought), since the horizontal resolutions of a 625-line 5·5 Mc/s picture and a 405-line 3 Mc/s picture are almost equal.

In September, 1955, the B.B.C. radio links between the cross-

Channel terminal near Dover and London were replaced by a coaxial-cable circuit installed by the Post Office. The B.B.C. and R.T.F. continued jointly to operate the cross-Channel radio links until July, 1959, when a permanent Post Office radio link was brought into service. The B.B.C. standards convertors are now installed in the Post Office terminal station at Tolsford Hill, near Dover.

Both the technical and the administrative co-ordination of these exchanges is undertaken by the E.B.U. During the past year, additional equipment has been installed at the E.B.U. technical co-ordination centre in Brussels to enable direct executive action to be taken there in switching certain Eurovision circuits. This has become necessary because of the increasing complexity of the vision network and of the lines that carry the accompanying sound. Members of the B.B.C. engineering staff are seconded to the E.B.U. for this work and others are lent for short periods from time to time.

The number of television programme exchanges in which the B.B.C. was involved during 1960 was

Incoming	259
Outgoing	117

By 1960, 18 television services in 14 countries had been linked together by the Eurovision network.

(16.3) Standards Convertors for 50/60 c/s

The exchange of live programmes with the American continent has yet to be achieved. The American television systems have a field frequency of 60 c/s, and the conversion system described above produces considerable flicker when the field frequencies differ by as much as 10 c/s. Means have been devised, however, for cancelling out this flicker,* and these have already been used for reproducing telerecordings on magnetic tape taken on the American standard and reproduced on the British standard, and vice versa.[46]

The method used is to insert a reference pulse at the end of each line before the signal is applied to the cathode-ray display tube, so as to produce a bright stripe on one side of the displayed picture. A pulse corresponding to this stripe appears in the output of the cameras, and this, like the picture information, is modulated at the difference between the field frequencies. This pulse is then selected by a gate and used to control the gain of an amplifier, so as to cancel out the flicker automatically.

An American company has now come to an agreement with the B.B.C. permitting it to manufacture, or have manufactured, standards conversion equipment in accordance with the B.B.C. design: the equipment is made by a British firm, which is also permitted by the B.B.C. to make and sell equipment.

(16.4) Cable Film

The use of the transatlantic telephone cable for the 2-way transmission of news films for television between Europe and America began in June, 1959, using a system and equipment developed by B.B.C. engineers.[47] In this system, known as Cablefilm,† the film is run at low speed through a film scanner and the vision signals so obtained are made to modulate a carrier of 5 kc/s which is transmitted over the cable. At the receiving end the film is reconstituted by a telerecording apparatus and shown at the normal speed.

Consideration of the characteristics of the cable indicated that a maximum video frequency of 4·5 kc/s could be used. It was therefore necessary to effect as many economies in the bandwidth of the video signal as are compatible with acceptab picture quality. These economies are as follows:

Restriction of the horizontal definition to that corresponding a bandwidth of 1·75 Mc/s in the 405-line television system.

A reduction to 200 lines using sequential scanning.

The scanning at the transmitting end of only alternate film fram with each frame-scan reproduced on two adjacent film frames the receiving terminal.

These measures result in reducing the 3 Mc/s bandwidth the British television system to approximately 450 kc/s, t remainder of the bandwidth reduction being obtained by decrease in the scanning speed until the maximum video fr quency corresponds to the available 4·5 kc/s upper limit. Th time required to scan the film is approximately 100 times norm and thus a half-minute news flash takes approximately 50 m to transmit.

Complete television programmes recorded as film are free exchanged internationally; this presents no major technic problem.

(17) RELAY EXCHANGES (WIRE BROADCASTING)

There are two commonly-used systems of distributing tel vision programmes by wire. In one of these, the two pr grammes (where available) are distributed at carrier frequenc Where the cable runs are short, as in blocks of flats or hote the Band I and Band III programmes may be distributed at the normal broadcast frequencies, but where large areas are to b covered it is usual to distribute the Band III programme o another Band I channel, using a frequency-changer at the mast distribution stations, in order to avoid the much greater cab losses which occur at frequencies in Band III. With this syste a viewer is able to choose a standard type of television receiver

In the alternative system, the programmes are distributed lower frequencies using multicore cables. Viewers' receive of a special design are supplied by the company operating th system, and a change of programme is effected by switching th receiver unit across the appropriate cable pairs. Alternativel adaptors are available which enable standard television receive to be used.

At the end of 1960, the total number of relay exchanges operation was 552, of which 64 were for television only, 281 f both sound and television and 207 for sound only. The tot number of subscribers was 1 074 432 including 450 725 televisio subscribers.

(18) ACKNOWLEDGMENTS

The author is indebted to the Director of Engineering of th B.B.C. for permission to publish the paper. Its preparation h been undertaken mainly by Messrs. H. T. Greatorex (Associa Member) and M. Campbell of the Engineering Informatio Department of the B.B.C. Many helpful comments have bee made by a number of other members of the Engineerin Division.

In conclusion the B.B.C. wishes to pay tribute to the technic achievements of the radio industry over the last 21 years, whic have made a major contribution to the progress described in th review. The B.B.C. also acknowledges the help and co-oper tion of the Post Office in the development of its plans in th technical as in other fields. Valuable work has been done b the several international organizations mentioned herein and al by the British Standards Institution and other profession bodies in this and other countries. Finally The Institutio itself has played an important part in maintaining profession standards, and in encouraging technical developments an making them known to engineers throughout the world.

* British Patent Application No. 3875, 1960.
British Patent No. 801140.
† British Patent Application Nos. 16323, 18277 and 18853, 1959.

(19) BIBLIOGRAPHY

(1) Ashbridge, Sir N.: 'Broadcasting and Television', *Journal I.E.E.*, 1939, **84**, p. 380.
(2) Television Committee 1943 Report (Hankey Committee). (H.M. Stationery Office.)
(3) 'Convention on the British Contribution to Television', *Proceedings I.E.E.*, 1952, **99**, Part IIIA.
(4) European Broadcasting Conference, Stockholm, 1952 (I.T.U., Geneva, June, 1952).
(5) Fowler, E. A.: 'The BBC Riverside Television Studios: The Architectural Aspects', BBC Engineering Division Monograph No. 13 (July, 1957).
(6) Nickels, H. C., and Grubb, D. M. B.: 'The BBC Riverside Television Studios: Some Aspects of Technical Planning and Equipment' (BBC Engineering Division Monograph No. 14, October, 1957).
(7) McGee, J. D., and Lubszynski, H. G.: 'E.M.I. Cathode-Ray Television Transmission Tubes', *Journal I.E.E.*, 1939, **84**, p. 468.
(8) Blumlein, A. D., Browne, C. O., Davis, N. E., and Green, E.: 'The Marconi–E.M.I. Television System', *ibid.*, 1938, **83**, p. 758.
(9) Polonski, J.: 'A French Portable Television Camera', *Journal of the Television Society*, April/June, 1958, **8**, p. 423.
(10) McGee, J. D.: 'A Review of Some Television Pick-up Tubes', *Proceedings I.E.E.*, Paper No. 931, April, 1950 (**97**, Part III, p. 377).
(11) Janes, R. B., Johnson, R. E., and Moore, R. S.: 'Development and Performance of Television Camera Tubes', *RCA Review*, 1949, **10**, p. 191.
(12) Janes, R. B., Johnson, R. E., and Handel, R. R.: 'A New Image Orthicon', *ibid.*, 1949, **10**, p. 586.
(13) Brothers, D. C.: 'The Testing and Operating of $4\frac{1}{2}''$ Image Orthicon Tubes', *Journal of the British Institution of Radio Engineers*, 1959, **19**, p. 777.
(14) Sproson, W. N.: 'The Measurement of the Performance of Lenses', *BBC Quarterly*, 1953, **8**, p. 17.
(15) Hacking, K.: 'An Optical Method for Obtaining the Frequency Response of a Lens', *Nature*, 1958, **181**, p. 1158.
(16) Sproson, W. N.: 'New Equipment and Methods for the Evaluation of the Performance of Lenses for Television' (BBC Engineering Division Monograph No. 15, December, 1957).
(17) Wood, C. B. B., Lord, A. V., Rout, E. R., and Vigurs, R. F.: 'The Suppressed Frame System of Telerecording' (BBC Engineering Division Monograph No. 1, June, 1955).
(18) Angel, Y: 'Considérations sur le fonctionnement des vidigraphes', *L'Onde Electrique*, 1954, **34**, p. 958.
(19) Pawley, E. L. E.: 'B.B.C. Sound Broadcasting 1939–60', *Proceedings I.E.E.*, Paper No. 3560, May, 1961 (**108** B, p. 279).
(20) Spooner, A. M., and Worswick, T.: 'Special Effects for Television Studio Productions', *ibid.*, Paper No. 1477, March, 1953 (**100**, Part I, p. 288).
(21) Chapman, N. F.: 'The Equipment of the BBC Television Film Studios at Ealing' (BBC Engineering Division Monograph No. 27, January, 1960).
(22) Kilvington, T., Laver, F. J. M., and Stanesby, H.: 'The London–Birmingham Television Cable System', *Proceedings I.E.E.*, Paper No. 1187, October, 1951 (**99**, Part I, p. 44).
(23) Faulkner, H.: 'Permanent Point-to-Point Links for Relaying Television', *ibid.*, Paper No. 1299, April, 1952 (**99**, Part IIIA, p. 313).
(24) Worswick, T., and Larkby, G. W. H.: 'An Improved Roving Eye' (BBC Engineering Division Monograph No. 12, April, 1957).
(25) Gallagher, J. C.: 'The Technical Problems of Broadcasting, Part 3—Linking the Television Network', *Engineering*, 29th January, 1960, p. 168.
(26) Quinton, E. C.: 'A UHF Television Link for Outside Broadcasts' (BBC Engineering Division Monograph No. 19, June, 1958).
(27) Bolt, F. D.: 'The Technical Problems of Broadcasting, Part 4—Feeding the Transmitting Aerial', *Engineering*, 12th February, 1960, p. 241.
(28) Cooper, V. J.: 'A Comparison between High-Level and Low-Level Modulation for Television Transmitters', *Marconi Review*, 1952, **15**, p. 118.
(29) Wharton, W., and Platts, G. G.: 'The Crystal Palace Band I Television Transmitting Aerial' (BBC Engineering Division Monograph No. 23, February, 1959).
(30) Monteath, G. D., Millard, G. H., and Whythe, D. J.: 'The Use of a High-Gain Television Transmitting Aerial in a Populous Area, with particular reference to the Crystal Palace Station', *Proceedings I.E.E.*, Paper No. 3334, January, 1961 (**108** B, p. 65).
(31) Bevan, P. A. T., and Page, H.: 'Sutton Coldfield Television Broadcasting Station', *ibid.*, Paper No. 1116, April, 1951 (**98**, Part III, p. 416).
(32) Weigall, D. B.: 'The Holme Moss Television Transmitting Station', *BBC Quarterly*, 1951, **6**, p. 182.
(33) Bevan, P. A. T.: 'Television Broadcasting Stations', *Proceedings I.E.E.*, Paper No. 1314, April, 1952 (**99**, Part IIIA, p. 179).
(34) Bevan, P. A. T.: 'The BBC Television Transmitting Stations', *BBC Quarterly*, 1952, **7**, p. 235.
(35) McLean, F. C., Thomas, A. N., and Rowden, R. A.: 'The Crystal Palace Television Transmitting Station', *Proceedings I.E.E.*, Paper No. 2069, March, 1956 (**103** B, p. 633).
(36) Rendall, A. R. A., and Padel, S. H.: 'The Broadcasting House—Crystal Palace Television Link', *ibid.*, Paper No. 2072, March, 1956 (**103** B, p. 644).
(37) Cooper, V. J., and Morcom, W. J.: 'Band I Television-Transmitter Design, with particular reference to the Transmitters at Crystal Palace', *ibid.*, Paper No. 2075, March, 1956 (**103** B, p. 651).
(38) Saxton, J. A., Harden, B. N.: 'Polarization Discrimination in V.H.F. Reception', *ibid.*, Paper No. 2184, November, 1956 (**103** B, p. 757).
(39) 'Television Field Trials of 405 Line and 625 Line Systems in the UHF and VHF Bands 1957–58' (BBC, May, 1960).
(40) Report of the Television Advisory Committee 1960 (H.M. Stationery Office, June, 1960).
(41) Gouriet, G. G.: 'An Introduction to Colour Television'. (Published for the Television Society by Norman Price, Ltd.)
(42) Sproson, W. N., Watson, S. N., and Campbell, M.: 'The BBC Colour Television Tests: An Appraisal of Results' (BBC Engineering Division Monograph No. 18, May, 1958).
(43) Atkins, I. R., Stanley, A. R., and Watson, S. N.: 'A Survey of the BBC Experimental Colour Transmissions' (BBC Engineering Division Monograph No. 32, October, 1960).
(44) Maurice, R. D. A.: 'The Technical Problems of Broadcasting: Part 5—Televising Internationally', *Engineering*, 26th February, 1960, p. 305.
(45) Lord, A. V.: 'Conversion of Television Standards', *BBC Quarterly*, 1953, **8**, p. 108.
(46) Lord, A. V.: 'A Standards Converter for Television Exchanges between Europe and North America', *EBU Review*, Part A, 1960, **63**, p. 209.
(47) Wood, C. B. B., and Shelley, I. J.: 'Cablefilm', *Journal I.E.E.*, 1960, **6**, p. 634.
(48) 'International Radio Regulations, Annexed to the International Telecommunication Convention, Atlantic City 1947 (H.M. Stationery Office, 1948).
(49) Weaver, L. E.: 'The Measurement of Random Noise in the Presence of a Television Signal' (BBC Engineering Division Monograph No. 24, March, 1959).

(20) APPENDICES

(20.1) B.B.C. Television Transmitting Stations

Station	Channel	Frequency		Effective radiated vision power	Polarization	Opening date
		Sound	Vision			
		Mc/s	Mc/s	kW		
Crystal Palace	1	41·50	45·00	200	V	28.3.56
Divis				12	H	21.7.55
Thrumster				0·25–7	V	15.12.58
Holme Moss	2	48·25	51·75	100	V	12.10.51
Dover				0·05–1·4*	V	21.4.58
North Hessary Tor				1–15*	V	17.12.54
Whitehawk Hill (Brighton)				0·4 max.*	V	5.8.59
Rosemarkie				1·5 max.*	H	16.8.57
Londonderry				1·5 max.*	H	18.12.57
Kirk o'Shotts	3	53·25	56·75	100	V	14.3.52
Rowridge				1–32*	V	12.11.54
Tacolneston (Norwich)				1·3–15*	H	1.2.55
Blaen-Plwyf				1–3*	H	29.4.57
Sutton Coldfield	4	58·25	61·75	100	V	17.12.49
Sandale				10–28*	H	5.11.56
Folkestone				0·007 max.*	H	14.7.58
Les Platons				1	H	3.10.55
Meldrum				4–17*	H	12.10.55
Wenvoe	5	63·25	66·75	100	V	15.8.52
Pontop Pike				12	H	1.5.53
Douglas, Isle of Man				0·18–2·8*	V	20.12.53
Orkney				4–14*	V	22.12.58
Peterborough				1	H	5.10.59

* Directional aerial.

The population coverage of the above stations is 98·8% of the population* of the United Kingdom (100 μV/m contour). It should be noted that there are a number of 'pockets' situated within the 100 μV/m contour where reception may be poor because of local screening; in some areas, reception may also be unsatisfactory at times because of co-channel or foreign interference.

(20.2) Proposed B.B.C. Television Satellite Transmitting Stations

Until sites have been chosen and technical operating conditions agreed, it is not possible to predict exactly the improvements in coverage which the above stations will achieve. It is estimated, however, that the 25 stations will increase the coverage by some 300 000 (0·6%) and give improved service for a further 1 400 000 people.

* The total population (1951 Census) was 50 369 000.

Stage I

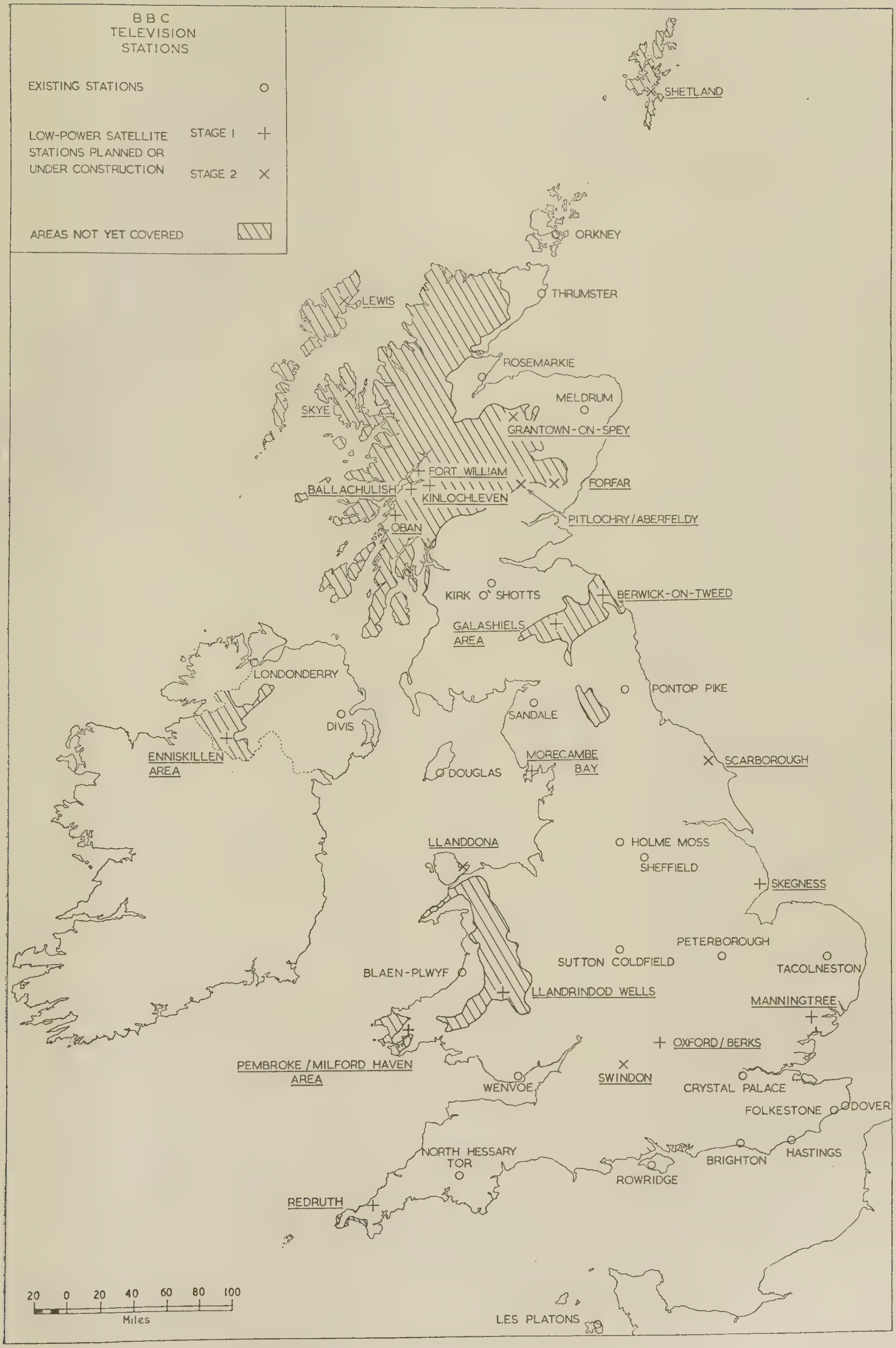
Fort William
Galashiels area
Berwick-on-Tweed
Llandrindod Wells area
Kinlochleven
Oban
Oxford
Redruth
Morecombe Bay
Enniskillen area
Manningtree
Pembroke/Milford Haven area
Sheffield
Skegness
Ballachulish

Stage II

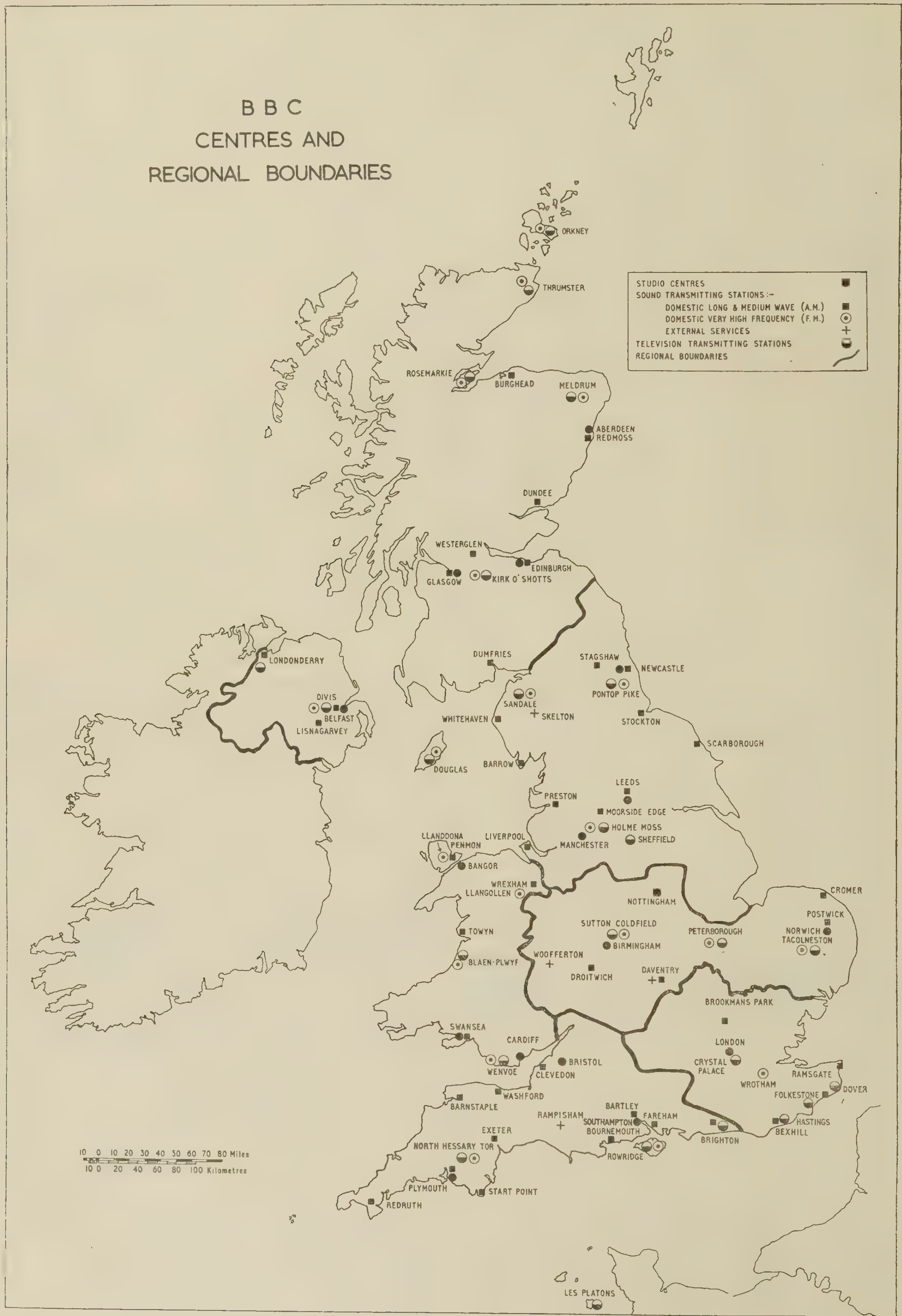
Forfar, Angus
Grantown-on-Spey
Lewis
Pitlochry/Aberfeldy
Shetland
Skye
Llanddona
Hastings
Scarborough
Swindon

Stage I is due to be completed during 1961 and 1962; construction of the stations in Stage II is proceeding concurrently and it is expected that most will be completed by the end of 1963 and the remainder in the spring of 1964.

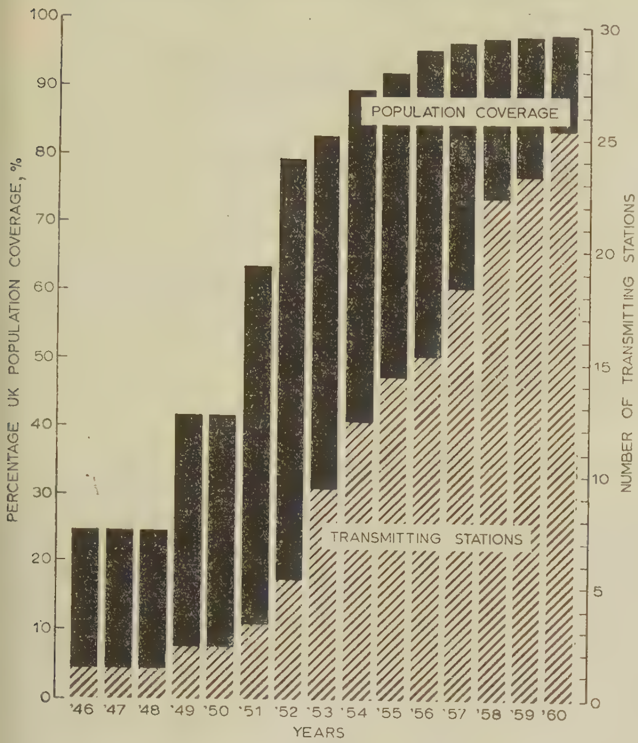
(20.3) B.B.C. Television Coverage



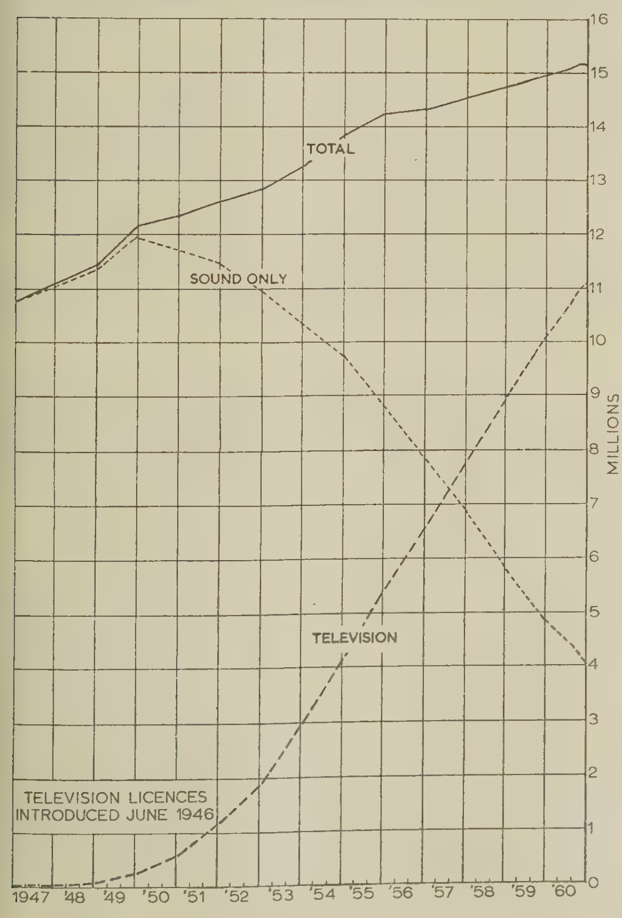
(20.4) Regional Boundaries, Transmitting Stations and Studio Centres



(20.5) **B.B.C. Television Population Coverage**



(20.6) **Growth of Broadcasting Receiving Licences**



(20.7) **Some Outstanding Dates in the Development of Television**

1st September, 1939	Television service closed prior to outbreak of war.
7th June, 1946	Television service reopened.
27th August, 1950	Broadcast of first television programme originating outside the United Kingdom. (Outside broadcast from Calais using B.B.C. equipment.)
21st May, 1950	Lime Grove television studio centre opened.
30th September, 1950	First 'live' television broadcast from an aircraft in flight.
8th July, 1952	First public transmission in the United Kingdom of a television programme from Paris using B.B.C. 819/405 line standards convertor).
2nd June, 1953	Coronation ceremony televised for the first time.
15th June, 1953	First 'live' television programme from a ship at sea during the Royal Naval Review.
27th January, 1954	B.B.C. Television Centre in London brought into use (Scenery Block).
6th June–4th July, 1954	First European exchange of television programmes (Eurovision) with eight countries taking part.
15th September, 1955	First section of 2-way permanent television link with Continent opened (London–Dover coaxial-cable circuit).
22nd September, 1955	Alternative television programmes began in the United Kingdom (I.T.A.).
10th October, 1955	B.B.C. began first series of experimental colour television transmissions from Alexandra Palace.
4th June, 1956	First programme from Riverside (Hammersmith) television studios.
16th June, 1956	First 'live' television broadcast from a submarine at sea.
4th August, 1956	First 'live' television broadcast from a helicopter in flight.
5th November, 1956	B.B.C. experimental colour television transmissions included 'live' studio productions for the first time.
24th September, 1957	B.B.C. television for schools began.
29th October, 1957	First B.B.C. unattended television studio brought into service (St. Stephen's Hall, opposite the House of Commons).
11th November, 1957	High-power experimental transmissions on 405 lines in Band V began from B.B.C. Crystal Palace station.
25th December, 1957	Her Majesty the Queen's Christmas broadcast televised for the first time.
5th May, 1958	High-power experimental transmissions on 625 lines in Band V began from B.B.C. Crystal Palace station.
28th October, 1958	State Opening of Parliament televised for the first time.
17th June, 1959	First public demonstration of B.B.C. Cablefilm system for transmitting news films for television over the transatlantic telephone cable.
24th June, 1959	B.B.C. satellite transmitting station plans announced (Stage I): 14 stations for television. [An additional station (Ballachulish) added 14th September, 1960.]
1st July, 1959	Second (cross-Channel) section of permanent 2-way television link with Continent completed.
19th December, 1959	New B.B.C. television standards convertor (European to N. American standards) used for first time to produce 525-line 60-field/sec videotape recordings of Heads of States Meeting in Paris.
20th May, 1960	Stage II of B.B.C. satellite transmitting station plans announced. Ten stations for television.
29th June, 1960	First programme from B.B.C. Television Centre (Studio 3).
27th August, 1960	First use of B.B.C. universal zoom camera (with zoom lens inside camera body).
8th September, 1960	Announcement by Postmaster General of Committee of Enquiry into the Future of Sound and Television Broadcasting (Pilkington Committee).

621.317.799.029.64 : 621.391.812.621

The Institution of Electrical Engineers
Paper No. 3553 E
July 1961

THE CONSTRUCTION AND PERFORMANCE OF AN AIRBORNE MICROWAVE REFRACTOMETER

By J. A. LANE, M.Sc., Associate Member, D. S. FROOME, B.Sc.(Eng.), Graduate, and G. J. McCONNELL.

(*The paper was first received* 9*th November*, 1960, *and in revised form* 19*th January*, 1961.)

SUMMARY

The paper describes the construction of a Birnbaum-type microwave refractometer and its application in studies of refractive-index variations at heights up to 20000ft. Practical problems encountered in airborne measurements are discussed in detail. Some typical results are given which illustrate the performance obtained in measurements of refractive-index profiles and discontinuities.

(1) INTRODUCTION

Studies of the influence of the medium between transmitter and receiver on wave propagation have always formed an important part of research in radiocommunication. Of particular interest at the present time are investigations of the physical structure of the lower atmosphere and its effect on long-distance transmission at frequencies above about 30Mc/s. The observed field strength at points beyond the horizon is tentatively explained in terms of a process of scattering or partial reflection due to spatial variations in the refractive index of the atmosphere between transmitter and receiver. Until recently, however, this approach has been hampered by the lack of adequate experimental data. The results obtained from radio-sonde ascents have insufficient detail, and the development of microwave refractometers for use in airborne measurements therefore represents an important advance in an experimental technique, having applications in both radio and meteorological research.

The two main types of refractometer developed to date are those due to Crain[1,2] and Birnbaum,[3,4] both methods making use of a resonant cavity as the essential feature of the measuring equipment. The techniques differ, however, in the methods used to measure and record the refractive-index fluctuations. In this paper, a Birnbaum-type instrument is described which has been operating recently in the United Kingdom in airborne investigations at heights up to 20000ft. It was designed primarily for studies of the refractive-index changes which occur over extended height intervals, i.e. profile measurements. It has also proved useful, however, in some horizontal soundings, e.g. in measurements of the relatively large variations observed in and near clouds at low altitudes. Some work has already been published in the United States describing the results of similar investigations,[2,4,5] but little information is at present available in published form on some aspects of refractometer design and performance. Some detailed comments on relevant practical problems are therefore given, together with a few typical results which illustrate the nature of the airborne experiments.

(2) PRINCIPLE OF METHOD

The fractional change, $\Delta f/f$, in the resonant frequency of a microwave cavity produced by a change Δn in the refractive index of the enclosed air is given (to a close approximation) by

$$\Delta n = -\Delta f/f \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

Δn is of the order of 10^{-6} for small fluctuations, and these changes are therefore defined in terms of N units, where

$$N = (n - 1)10^6 \quad . \quad . \quad . \quad . \quad . \quad (2)$$

We have, therefore,

$$\Delta N = -10^6\Delta f/f \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Microwave refractometers record the variation in the parameter N by measuring Δf; a change of $10N$ units ($\Delta N = 10$) in a resonant cavity operating at 10Gc/s thus corresponds to a change of 100kc/s in the resonant frequency. It is at once apparent, therefore, that a satisfactory measurement of small refractive-index changes imposes severe requirements on the stability, noise level and calibration of microwave refractometers.

The principle of the Birnbaum instrument is illustrated in Fig. 1. A frequency-modulated reflex klystron is connected via a waveguide T-junction to a sealed reference cavity and a sampling or test cavity open to the atmosphere. The resonant frequencies of the two cavities differ by approximately 4Mc/s and oscillations are generated in each cavity in turn. Rectified voltages derived from crystal detectors attached to the cavities produce two trigger pulses which operate a bistable multivibrator circuit. Since the pulses are displaced with respect to each other in time, the width of the square wave at the multivibrator output is proportional to the difference between the resonant frequencies of the cavities. The average voltage of the square wave is applied to a recorder or meter calibrated in N units. The method is, in effect, a technique for 'sampling' the resonant frequency of the test cavity at time intervals of t seconds where $1/t$ is the modulation frequency in cycles per second.

The main features of the equipment are described in the next Section, details of design being omitted where these are similar to those already given in published work.

(3) DESCRIPTION OF EQUIPMENT

The main units in the complete refractometer consist of (*a*) electronic equipment, including modulator, pulse-shaping circuits, power supplies, etc., (*b*) resonant cavities and associated waveguide components, and (*c*) recording equipment.

(3.1) Modulator and Pulse-Shaping Circuits

The modulating voltage, having a saw-tooth waveform, is produced by a bootstrap sweep circuit[6] with a thyratron as the switching tube. The modulation frequency is 100c/s, and the saw-tooth waveform is synchronized to the 50c/s input supply. The output of the klystron is fed via a matched E-plane T-junction to the two cavities, and small vane-type attenuators (of about 15dB attenuation) are included in the waveguide connection following the junction to prevent frequency pulling of the klystron and to minimize the effect of coupling between the high-Q cavities. A ferrite attenuator (isolator) is also included in the waveguide connection to the sampling cavity. The importance of these attenuators, especially in airborne work, is discussed in Section 8.

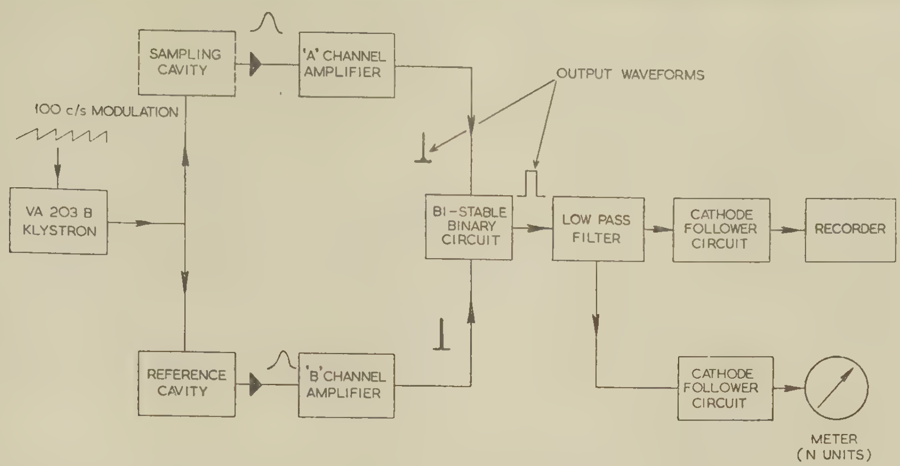


Fig. 1.—Schematic of Birnbaum-type refractometer.

The remainder of the electronic equipment consists of two identical amplifier and pulse-shaping circuits, a bistable multivibrator, a filter section, and the output circuit. The *RC*-coupled amplifiers have a gain of about 72 dB, and amplify the voltage waveform derived from the crystal detectors without appreciable distortion. After amplification of the resonance pulse, a limiting circuit is arranged to pass the peak (the top 10%) of the input waveform; this is then amplified and differentiated, with the negative pulse by-passed to earth via a germanium diode. The waveform at this stage is a positive 20 V pulse, the amplitude of which is largely independent of variations in the input voltage to the *RC*-coupled amplifiers. A positive square wave, of 50 V amplitude, is obtained from the multivibrator and a clamping circuit. As explained above, the precise duration of the square wave depends on the relative time displacement of the two trigger pulses in the two amplifier channels. The klystron mode curve, the cavity responses, the trigger pulses and the final square wave can all be observed as required on a small cathode-ray oscillograph included in the refractometer.

(3.2) Output Circuit

The output circuit differs slightly from the circuit described by Birnbaum,[3] several modifications being introduced to facilitate the use of the refractometer in airborne measurements.

The filter circuit following the multivibrator section consists of a twin-T rejection filter (tuned to 100 c/s) and a low-pass filter. The latter produces an average voltage from the square wave and also attenuates the 100 c/s component and the harmonic frequencies. After initial adjustment of the circuit components, the square wave (50 V amplitude) occupies 2/25th of the modulation period, i.e. 8×10^{-4} sec. The amplitude of the saw-tooth modulation is such that the mean width of the square wave is produced when the resonant frequencies of the cavities differ by 4 Mc/s; in this condition, the mean output voltage is 4 V. A 'backing-off' voltage is provided in the output stage so that the voltage input to the recorder is then zero. From eqn. (3) it can be seen that, at the operating frequency of 9 450 Mc/s, a change of refractive index of 10 N units in the sampling cavity produces a change, Δf, of 94 kc/s in the resonant frequency; this in turn produces a change of 0·094 V in the average value of the square wave. The resistance network in the output circuit is so constructed that changes of refractive index of ± 50, 100 or 200 N units produce full-scale deflection (± 2 in) of a galvanometer trace on photographic paper. A further switch is provided which changes, in small increments, the backing-off voltage mentioned above; this has the effect of adding or subtracting a voltage in the galvanometer circuit equivalent to a change of 50, 100, 150 or 200 N units in the refractive index. This facility is especially valuable in airborne measurements in that a total change of, say, 200 N units can be recorded at maximum sensitivity, i.e. a change of $\pm 50\,N$ units giving ± 2 in deflection. Unfortunately, the galvanometer deflection cannot be observed whilst recording is in progress, and an additional output circuit, similar to the above, is included, as shown in Fig. 1. This circuit operates a panel-type microammeter calibrated in N units, and synchronous switching of recorder and meter ensures that the recorder trace is always kept on the photographic chart. If desired, recordings of refractive-index changes can be made at ground level on a pen recorder, in which case full-scale deflections equivalent to changes of ± 10, 20, 50 or 100 N units can be obtained. The frequency response of the output circuit and the relatively low chart speed generally used in airborne measurements impose an upper limit of about 3 c/s on the frequency of fluctuations which can be recorded.

No special attempt was made in developing the instrument to reduce the size and weight to a minimum, and the first model described above consists of a main refractometer assembly and a power supply, each having dimensions of approximately 16 in × 8 in × 10 in. These two units together weigh about 65 lb.

(3.3) Resonant Cavities

The sampling and test cavities were designed for use in the region of 9 400 Mc/s using the H_{011} mode, with the axial length made equal to the diameter to ensure a maximum Q-factor.[7] Both cylinders were machined from Invar and then silver-plated. The reference cavity, located inside the main refractometer unit, can be tuned over a small range by means of an Invar screw moving through an O-ring seal. In the early stages of the investigation, it was thought doubtful whether a tunable, evacuated cavity could be made which would be suitable for airborne use over a long period. For this reason, a sealed air-filled cavity was preferred. Open end-plates of the type shown in Fig 2 were fixed to the sampling cavity, about 60–70% of the end-plate being removed to facilitate the free passage of air through the cavity. An annular ring remains in the region of maximum circumferential current flow. The loaded Q-factor of the sampling cavity is approximately 7 000, compared with a value of 10 000 for the reference cavity. The coupling holes are 0·25 in in diameter and, in the reference cavity, are sealed with mica covered with Araldite.

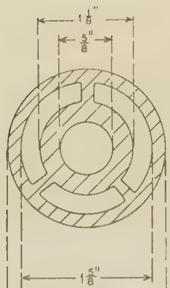


Fig. 2.—End-plates of 3 cm-band sampling cavity.

(3.4) Calibration of Refractometer

A calibration of the equipment requires a technique for simulating known, small changes in the resonant frequency of the sampling cavity. This was effected by replacing the sampling cavity by an H_{011}-cavity wavemeter specially constructed for the purpose. The tuning plunger was a 0·25 in rod operated by a micrometer drum 2·5 in in diameter. The wavemeter was in turn calibrated against a reference standard using the comparison technique described by Sproull and Linder.[8] This simple but effective method enables small changes in the resonant frequency of high-Q cavities to be measured with an uncertainty of about 1 part in 10^6. Known changes of resonant frequency, equivalent to changes of 50, 100 or 200 N units in refractive index, were produced by the calibrated wavemeter, and the resulting galvanometer deflection in the recorder was then adjusted to the desired value by a small change in the amplitude of the modulating voltage.

Even with Invar cavities, the apparent change in measured refractive index due to the effect of temperature changes on cavity dimensions is not negligible in measurements over height intervals of several thousand feet. Some additional experiments, to investigate this effect and other features of interest, were therefore carried out in which the sampling cavity was located in an open wind tunnel in an air stream of known velocity, temperature and humidity. Changes in the air temperature could be made only relatively slowly, and this restriction imposed some limitation on the accuracy of the results because of instrumental drift during the experiment; nevertheless, a few useful results were obtained which indicated an expansion coefficient of about 1·2 parts in 10^6 per deg C. It also appeared that no significant change in measured refractive index is produced by changes in air speed through the cavity. at least at speeds up to 250 m.p.h. A stream of air containing fine water droplets was also directed at the cavity; but with the end-plates or the type described, the danger of spurious results due to the collection of water inside the cavity appears to be negligible. It was, however, thought worth while to install a sealed window in the waveguide connection to the sampling cavity in order to prevent a stream of cold air from passing into the main assembly containing the klystron and reference cavity. The temperature variations in the latter cavity during subsequent airborne measurements were generally not greater than one or two degrees Celsius.

Recent work in the United States has illustrated the difficulties involved in attempts to obtain temperature-compensated cavities by inserting end-plates constructed partly from steel,[9] or by constructing the sampling and reference cavities from the same block of metal. In any case, the accuracy of measurement of relatively large refractive-index eddies and discontinuities extending over, say, a few hundred feet is not appreciably affected by the dimensional changes mentioned above. For these reasons the sampling cavity was not modified further; corrections are therefore required in the subsequent analysis of profile measurements over extended height intervals. This requirement is discussed again in Section 4.2.

(4) AIRBORNE INVESTIGATIONS

With the co-operation of the Meteorological Research Flight, Air Ministry, and the Royal Aircraft Establishment, Ministry of Aviation, arrangements were made to install the equipment in a Hastings transport-type aircraft used in meteorological research.

(4.1) Installation and Performance

The sampling cavity is located in the position shown in Fig. 3, approximately 3 ft outside the fuselage and about 15 ft in front

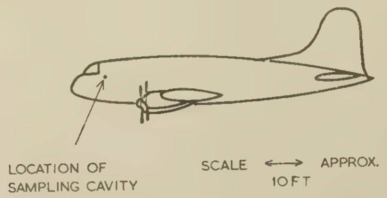


Fig. 3.—Location of sampling cavity of refractometer in Hastings aircraft.

of the inner port engine. The cavity and its connecting waveguide are supported by means of a drawn-steel tube which slides on a metal ramp attached to a bulkhead immediately behind the co-pilot's seat, and by this means the cavity can be locked in position or withdrawn into the aircraft as required.

The remainder of the equipment is installed in the main cabin, and with this arrangement a 25 ft length of waveguide is required between the sampling cavity and the main assembly. Although necessary in the circumstances described, such a disposition of equipment is not an ideal one, and for reasons discussed below, a more compact arrangement would probably have resulted in improved performance.

During the first flights made with the refractometer, it was evident that the noise level was appreciably greater than that experienced in the laboratory. This effect was initially attributed entirely to vibration of the equipment, which was occasionally severe, especially in turbulent conditions. Information was therefore obtained on the amplitude and frequency of vibration at several locations in a Hastings aircraft during a climb, a descent and in level flight. The results were then used in a series of laboratory tests in which the several units of the refractometer were attached to a vibration table, operating in the frequency range 0–100 c/s, and then subjected to horizontal and vertical displacements up to a maximum amplitude of about $\pm 0{\cdot}05$ in. The results indicated that sections of flexible waveguide in the connection to the sampling cavity were of doubtful value; when vibration effects were severe, flexing of these sections appeared to produce spurious fluctuations in the output equivalent to one or two N units, presumably as a result of the coupling between mismatched impedances (see Section 8). The results also demonstrated the marked superiority of air-damped anti-vibration mountings over the radial-spring type used previously. It was obvious, however, that vibration effects, although important, were not solely responsible for the relatively high noise level, which in the early stages of the work occasionally reached 2 N units (peak-to-peak amplitude). Some further improvement was obtained, following the vibration tests, by installing new electrical generators, but a residual noise level remains, in the model described above, which has a maximum amplitude of approximately 1 N unit (p–p). Microphonic effects

in the amplifiers, electrical pick-up in the output circuit and instability in the modulating voltage are thought to be the main causes. The long-term stability, especially important in profile measurements, showed a considerable improvement when the original reflex klystron (type 723 A/B) was replaced by a modern ruggedized klystron (type VA 203B).

Subsequent flights, carried out in the summer of 1959, were made to determine the performance of the modified equipment under a variety of conditions. Arrangements were made to note additional meteorological data (temperature, humidity, air speed, etc.), and a coded signal was inserted manually on a spare galvanometer trace on the recorder chart to give height markings at regular intervals. These simple techniques proved adequate in the preliminary investigations, but are now being replaced by the automatic recording of as many relevant parameters as possible.

(4.2) Results and Discussion

The flight procedure followed in obtaining the refractive-index data varies in detail according to the particular type of investigation being carried out. In general, initial connections and adjustments are made immediately after take-off, following, whenever possible, a preliminary operating period of 10 to 15 min on the ground during which the equipment reaches a stable condition. The refractometer is then set to the range required (usually the 50 N range), with the 'add N units' control in such a position that the recorder trace can be maintained on the chart throughout the remainder of the flight.

A typical sounding may consist of horizontal flights at several predetermined altitudes, a descent at, say, 2 000 ft/min over an extended height interval, or a 'saw-tooth' ascent and descent between given heights extending over several tens of miles in the horizontal direction. This last type of flight is especially valuable in studies of the nature of extended refractive-index discontinuities having a layer-type structure.

A typical profile measurement, replotted from the original photographic chart, is shown in Fig. 4. Absolute values of

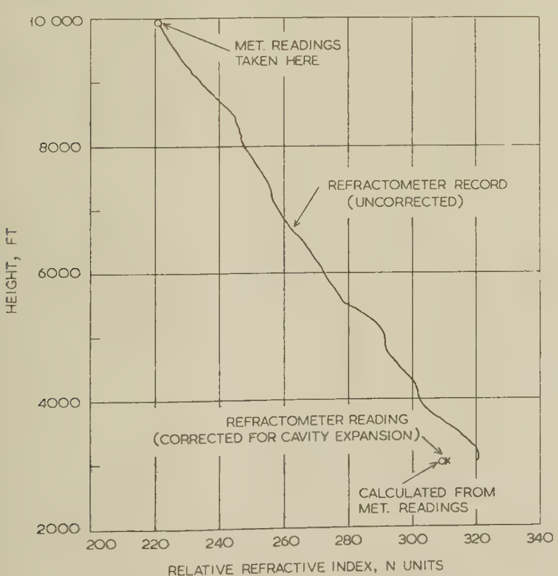


Fig. 4.—Comparison of refractometer record with meteorological data in a profile measurement.

refractive index can be assigned to such profiles by means of calculations based on meteorological observations taken at some convenient point during a flight, e.g. in clear non-turbulent air at, say, 10 000 ft. When conditions are suitable, these observations are made at several heights. Following this procedure, a comparison can be made between the change in refractive index measured by the refractometer and that calculated from the meteorological data. Points illustrating this comparison process are plotted in Fig. 4. The first datum point was obtained at 10 000 ft, and the difference between the refractive index values given by the two methods at 3 000 ft is almost exactly that expected assuming an expansion coefficient of about 1·2 parts in 10^6 per deg C in the sampling cavity. This result may be regarded as providing indirect confirmation of the value obtained in the wind-tunnel experiments already described; furthermore, it confirms the accuracy of the refractometer when used to record the vertical gradient of refractive index.

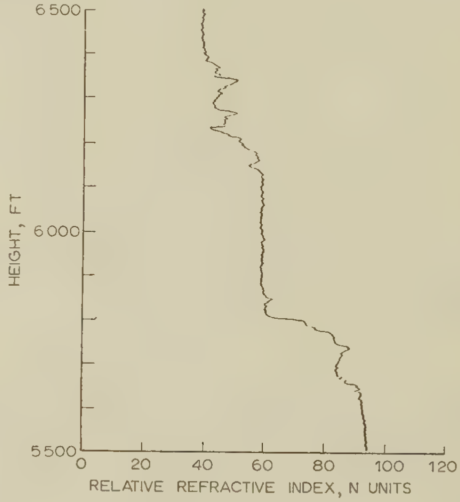


Fig. 5.—Specimen record from microwave refractometer showing discontinuities in vertical gradient of refractive index.

Southern England, 15th September, 1959.

A portion of a profile measurement is shown in detail in Fig. 5, and illustrates the nature of the trace obtained when the aircraft passes through a region having a marked discontinuity in the vertical gradient of refractive index. In this example, a change of about 20 N units occurs in a vertical interval of 50 ft. The flights carried out during the prolonged periods of anticyclonic weather during the summer of 1959 yielded many examples of discontinuities of this nature. Fig. 6 shows some refractive-

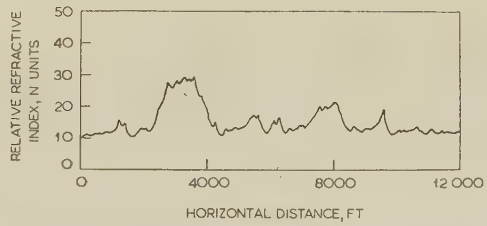


Fig. 6.—Variation in refractive index in cloud base at height of 3 000 ft.

Southern England, 25th June, 1959: 1 400 h G.M.T.

index changes observed in a level flight at 3 000 ft through a thin cloud base. Variations as large as 5–15 N units are apparent over distances of 200–2 000 ft.

The larger variations in the above results are mainly due to local humidity variations, and it is therefore of interest to cal-

culate the sensitivity of a microwave refractometer as a recorder of humidity fluctuations. We may express the refractive index as

$$(n-1)10^6 = N = 77{\cdot}6\frac{P}{t} + 3{\cdot}735 \times 10^5\frac{P_w}{t^2} \quad . \quad . \quad (4)$$

where p = Total pressure, mb.
p_w = Water-vapour pressure, mb.
t = Temperature, deg K.

The small variations in total pressure over small distances result in negligible variations in N; therefore, by taking the total differential of this expression and putting $dp = 0$, we obtain an expression of the form

$$dN = k_1 dp_w - k_2 dt \quad . \quad . \quad . \quad . \quad . \quad (5)$$

in which k_1 is a function of t, and k_2 is a function of p, t and p_w. If dN, dp_w and dt represent the amplitudes of the corresponding small fluctuations, k_1 and k_2 may be found for various mean conditions. Typical values are:

Ground level (288° K): $k_1 = 4{\cdot}5$, $k_2 = 1{\cdot}4$.
At 10 000 ft (268° K): $k_1 = 5{\cdot}2$, $k_2 = 0{\cdot}9$.

Assuming a relatively small contribution to dN as a result of temperature fluctuations, a change of, say, 5 N units corresponds to a change in partial pressure of about 1 mb, i.e. a change of approximately 6% in relative humidity at 288° K. Greater sensitivity in humidity measurement is possible, of course, with improved refractometers (with about 1 N unit full-scale deflection) and simultaneous temperature recordings.

(5) CONCLUSIONS

The results given above provide an indication of the type of information obtained from airborne measurements of refractive-index fluctuations using a microwave refractometer. In the development of these instruments, there are many practical problems which require detailed attention, especially in airborne applications. Nevertheless, the technique is an extremely valuable one and yields detailed information on the structure of the atmosphere, which is important in both radiocommunication and meteorology. In measurements of the vertical gradient of refractive index there is some uncertainty in the precise interpretation of a given profile owing to the relatively large horizontal distance traversed by the aircraft. However, this uncertainty is considerably reduced in 'saw-tooth-type' soundings, which are particularly valuable in studies of inversions and layer discontinuities.

The addition of absolute values to the refractometer record, in the manner described, also enables the information obtained to be used in studies of climatology.

(6) ACKNOWLEDGMENTS

The airborne experiments were planned in co-operation with Dr. R. J. Murgatroyd, Meteorological Research Flight, and the authors are especially indebted to Mr. D. D. Clark for invaluable help in organizing the experimental facilities. The installation of the airborne equipment was carried out at the Royal Aircraft Establishment, and the original Invar cavities were constructed at the Royal Ordnance Factory, Maltby. The help given by Dr. J. S. McPetrie in these arrangements is also gratefully acknowledged. At the Radio Research Station, Dr. J. A. Saxton provided general advice throughout the work. Mr. B. N. Harden was responsible for the experimental work in the early stages of the construction of the refractometer.

The work described was carried out as part of the programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(7) REFERENCES

(1) Crain, C. M., and Deam, A. P.: 'An Airborne Microwave Refractometer', *Review of Scientific Instruments*, 1952, **23**, p. 149.
(2) Crain, C. M.: 'Survey of Airborne Microwave Refractometer Measurements', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 1405.
(3) Birnbaum, G.: 'A Recording Microwave Refractometer', *Review of Scientific Instruments*, 1950, **21**, p. 169.
(4) Bussey, H. E., and Birnbaum, G.: 'Measurement of Variations in Atmospheric Refractive Index with an Airborne Microwave Refractometer', *Journal of Research of the National Bureau of Standards*, 1953, **51**, p. 171.
(5) Ament, W. S.: 'Airborne Radiometeorological Research', *Proceedings of the Institute of Radio Engineers*, 1959, **47**, p. 756.
(6) Terman, F. E.: 'Electronic and Radio Engineering' (McGraw-Hill, 1955), p. 651.
(7) Montgomery, C. G.: 'Technique of Microwave Measurements', M.I.T. Radiation Laboratory Series (McGraw-Hill, 1947), p. 301.
(8) Sproull, R. L., and Linder, E. G.: 'Resonant Cavity Measurements', *Proceedings of the Institute of Radio Engineers*, 1946, **34**, p. 305.
(9) Crain, C. M., and Williams, C. E.: 'Method of Obtaining Pressure and Temperature Insensitive Microwave Cavity Resonators', *Review of Scientific Instruments*, 1957, **28**, p. 620.
(10) Montgomery, C. G.: *loc. cit.*, p. 311.

(8) APPENDIX

(8.1) Resonant Frequency of Cavity in a Transmission Line

Consider the resonant frequency of a cavity located in a waveguide between a generator and load. If the generator and load are not matched to the connecting waveguide transmission line, the frequency at which maximum transmission occurs depends on the generator and load impedances, and on the phase of the voltage standing-wave ratios at a given reference plane.

If the voltage standing-wave ratios (V_{max}/V_{min}) resulting from mismatches at generator and load are r_1 and r_2, respectively, the difference, Δf, between the frequency for maximum transmission under these conditions and that with matched generator and load is given by[10]

$$\Delta f/f = \pm \left[\beta_1\left(r_1 - \frac{1}{r_1}\right) + \beta_2\left(r_2 - \frac{1}{r_2}\right)\right]\Big/4Q_0 \quad . \quad (1)$$

where Q_0 is the unloaded Q-factor of the cavity, and β_1 and β_2 are input and output coupling parameters determined from impedance measurements with a resonant cavity. Suppose the load is matched ($r_2 = 1$), and the unloaded Q-factor is 15 000; then with typical values of $\beta_1 = 0{\cdot}5$ and $r_1 = 1{\cdot}1$, we have, for $f = 9\,500$ Mc/s,

$$\Delta f = \pm\, 0{\cdot}016\ \text{Mc/s}$$

This is equivalent to a change of $\pm 1{\cdot}7\,N$ units, and corresponds to the maximum possible change in resonant frequency in the above example. Vibration effects, even with flexible waveguide, would not normally produce such a large deviation, but the calculation serves to show the order of magnitude of the likely variation.

In addition, the phase of the source impedance, as seen at the input to the sampling cavity, will vary considerably in the frequency range over which the modulated klystron is oscillating. It is important, therefore, to ensure that both cavities in the refractometer are located, so far as possible, between fixed matched impedances.

621.387.464 : 550.378

The Institution of Electrical Engineers
Paper No. 3545
July 1961

THE AERIAL SURVEY OF TERRESTRIAL RADIOACTIVITY

By D. WILLIAMS, B.Sc., A.Inst.P., and H. BISBY, B.Sc., Associate Member.

(*The paper was first received* 23*rd September*, 1960, *and in revised form* 13*th January*, 1961.)

SUMMARY

The basic methods of aerial radiometric survey are described and the factors determining their sensitivity, precision and utilization are discussed. Particular reference is made to the problem of discriminating between radiations from the earth's surface and those from other sources. It is demonstrated that the technique is of value in uranium exploration and for surveying deposited radioactive contamination.

LIST OF SYMBOLS

$B(h)$ = Build-up factor.
h = Height of detector above ground surface.
ϕ_0 = γ-flux at ground level.
ϕ_h = γ-flux at height h.
k = Constant.
n = Total mean counting rate.
n_d = Datum mean counting rate.
n_T = Terrestrial γ-flux mean counting rate.
t = Integration time of instrument for survey.
t_d = Integration time of instrument for datum determination.
v = Aircraft ground speed.
τ_1 = Time-constant of instrument for optimum point-source detection.
τ_2 = Time-constant of instrument for absolute measurements.
μ = Total linear absorption coefficient of air.

(1) INTRODUCTION

Surveys of the γ-radiation emitted by natural radioactive materials contained within the earth's surface have been of assistance in the search for new sources of uranium. High intensities of γ-radiation occur locally over surface exposures of uranium, thorium and potassium 40 minerals, and since all rocks are to some extent radioactive, γ-intensity patterns can be correlated with surface geology as an aid to geological mapping. The instruments and techniques used in geological radiometric surveys can be applied to the detection of radio-elements deposited on the earth's surface, provided that the γ-emissions have suitable energy and intensity and information is available of the pre-deposition γ-intensity patterns of the survey area. Such surveys can be of value in determining the extent of contamination in the environs of a nuclear installation should an abnormal release of radioactive material occur.

In both geological and 'environmental' applications, large areas of terrain must be surveyed, so that survey methods must be chosen for speed and efficiency. The methods fall broadly into two categories, i.e. car-borne and airborne surveys, in which the detector is traversed along predetermined routes at a fixed height above ground (about 6 ft and 500 ft respectively), and the radiometric record is correlated with detector plan-position. Both survey methods have advantages and disadvantages in a particular application, and are complementary rather than alternative techniques. The paper deals primarily with aerial survey and the factors determining its sensitivity, precision and utilization.

The object of aerial radiometric surveys is therefore to measure the intensity of terrestrial γ-radiation and its variation with position of the detector over the earth's surface. However, the indication of a γ-detecting instrument has additional contributions from γ-emitters in the atmosphere, together with cosmic radiation and the inherent radiation of the detector. Some of these contributions can vary independently in time and with position and complicate the problem of observing the contribution from the ground.

(2) γ-RADIATION FROM TERRESTRIAL RADIOACTIVITY

(2.1) Natural Radioactivity

The γ-emissions from uranium minerals arise mainly from the decay of lead 214 and bismuth 214, usually in or near secular equilibrium with the parent element uranium 238, and comprise a number of 'primary' radiations at discrete energies in the range 0–2·4 MeV. Similarly, the primary γ-radiations from thorium minerals are in the range 0–2·7 MeV, largely owing to the decay of thorium 232 daughter elements actinium 228, bismuth 212 and thallium 208. Potassium 40, the third natural source of γ-radiation has a simple decay, emitting primary radiation at 1·48 MeV only.

In nature the primary emissions are modified by absorption and scattering in the source material, in the nearby rock and in the air between the source and the detector. The radiation reaching the detector is thus composed of the original primary emissions, with reduced intensity, superimposed on an energy continuum of 'secondary' degraded radiation, and has its origin in a surface layer of rock less than 30 cm thick. The intensity of this radiation is normally constant in time at a given place, but varies with position according to surface geology. High intensities occur locally over radiactive minerals and contrast with lower intensities over the country rocks, of which granites are more radioactive than many sedimentary formations such as limestones. Very low intensities occur over stretches of water which are deep enough to absorb all the radiation from the underwater rocks and of sufficient extent to exclude that from neighbouring land.

(2.2) Deposited Radioactivity

The earth's surface can be contaminated with fission and irradiation products from the explosion of nuclear weapons or by the effluents from nuclear installations. Deposited radioactivity from normal effluents is of low concentration, and is best monitored by methods other than aerial survey, but abnormal conditions might result in the deposition of activity over a wide area and would require an effective and rapid survey. The contaminant most likely to constitute an immediate health hazard is iodine 131 (mean γ-energy of 0·4 MeV and half-life of 8 days), since its deposition on herbage can be followed by its appearance in milk within 24 hours. The Medical Research Council's recommended maximum permissible level (m.p.l.) of iodine 131 in milk is 0·065 μc/l, and the equivalent m.p.l. level of contamination on herbage has been estimated[1] at 0·5 $\mu c/m^2$. Additional γ-emitters may be deposited whose relative pro-

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Williams and Mr. Bisby are at the U.K.A.E.A. Atomic Energy Research Establishment, Harwell.

portions depend on the irradiation history of the material prior to release of activity, but the detection of the m.p.l. of iodine 131 alone imposes the most stringent requirements on the survey technique. Since the emitted radiations are of lower energy than those of natural terrestrial activity, they suffer greater attenuation in the air and the γ-intensity produced by the m.p.l. of iodine 131 is comparable with the variations arising from the positional differences in the natural activity of surface rock.[2] Thus, unless the radiation from deposited contamination can be distinguished from the natural radiation of surface rocks, the pre-incident γ-intensities must be accurately known if post-incident iodine 131 contamination, at the m.p.l., is to be determined.

(2.3) Variation of Terrestrial γ-Flux with Height above the Earth's Surface

The γ-flux from terrestrial activity varies with height above ground and is determined by the nature of the source, its surface area, its thickness and the intervening air distance and density. For a source whose surface dimensions are much smaller than h, the γ-flux is given by

$$\phi_h = \phi_0 \varepsilon^{-\mu h} \mathrm{B}(h)/h^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

B(h) takes into account radiation scatter by the air; it is unity when h is zero and increases with h through the increasing probability of multiple Compton scattering. This has been expressed empirically as a power series of μh, and for single-medium problems, build-up effects are calculable[4, 5] by numerical methods. However, factors such as scattering in deep sources and backscattering of radiation from deposited contamination by the ground, complex energy spectra and the energy-dependence of detectors, make a semi-empirical approach based on direct experiment more profitable.[6, 7]

As the surface dimensions of the source increase, eqn. (1) progressively modifies to the case for an infinite source, when air absorption and scatter entirely account for the flux variations with height, namely

$$\phi_h = k\phi_0 \varepsilon^{-\mu h} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where k is a constant.

Fig. 1 shows an experimental evaluation for an infinite source (uniformly radioactive surface rock) and contrasts the decrease of γ-flux with the more rapid reduction over a radium point source. Similar, but more severe, attenuation with height occurs over point and infinite plane iodine 131 sources. The reduction of γ-flux with increasing h is not serious for the survey of large-area sources as encountered in geological mapping and environmental survey, but the detection of a point source is made difficult by the large reduction in the ratio of its signal to that of the country rocks. In practice, point sources are seldom encountered, and the reduction indicated above is less marked for more typical small-area (20 ft diameter) sources, since, until $h > 20$ ft, the attenuation follows the large-area-source case.

One of the main advantages of aerial survey over car-borne survey lies in the improved ground cover obtained. Fig. 2 shows that, over an infinite plane source, 50% of the radiation detected at $h = 10$ ft originates from a strip of ground 160 ft wide, whereas at $h = 500$ ft the width increases to 1 280 ft. Thus, the same effectiveness of sampling over a flat area can be obtained with a wider spacing of traverse lines if the height is increased, while over uneven ground there is the additional advantage that local screening effects can be avoided. Furthermore, the rate of change of flux with horizontal displacement from a point source decreases as h increases, and thus, at greater ground clearances, deviations from the planned route are less important and the speed of traverse can be increased without calling for faster response by the instruments. An upper limit to detector ground clearance is, however, imposed by air attenuation.

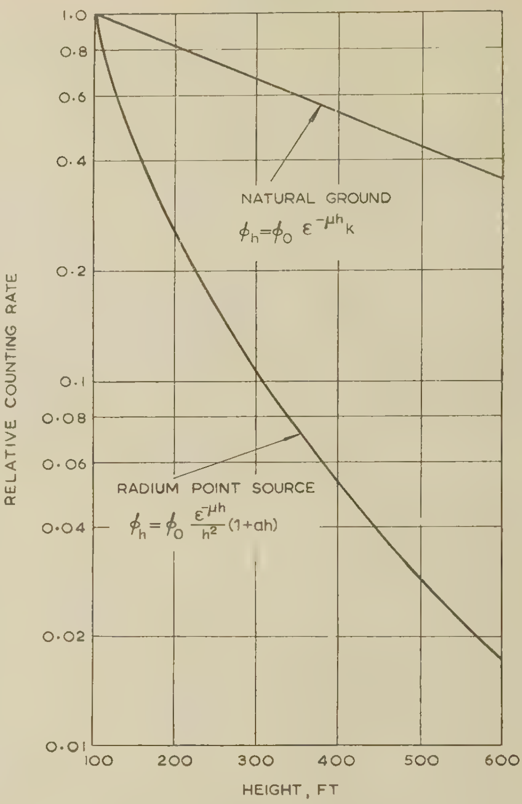


Fig. 1.—Variation of counting-rate with detector height above an infinite plane of uniformly radioactive surface rock and a radium point source.

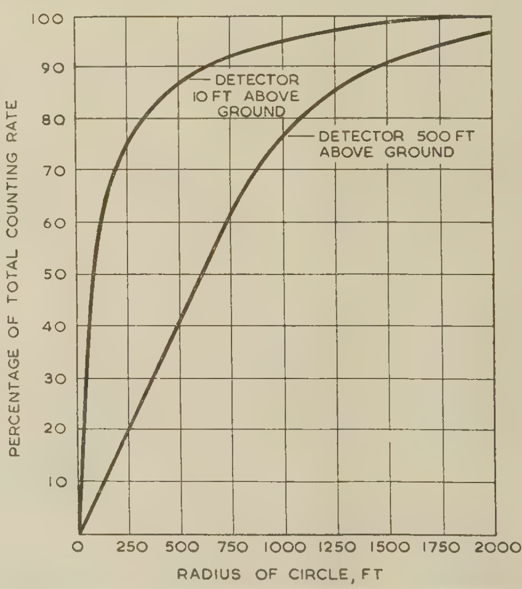


Fig. 2.—Detector at 10 ft and 500 ft above an extensive area of uniformly radioactive surface rock.
Proportion of total counting-rate contributed from within circles of radius r.

(3) OTHER SOURCES OF RADIATION

In addition to γ-radiation from terrestrial sources, there are contributions to the total counting rate observed from cosmic

radiation, γ-radiation from radio-elements in the atmosphere and radiation from radio-elements in the detector and associated apparatus. The sum of these contributions forms a datum of mean counting rate which must be subtracted from the total mean counting rate to obtain the terrestrial radiation contribution. Since the counts occur randomly in time, the accuracy of determining n_T is given by the relative standard deviation

$$\frac{\left(\frac{n}{t} + \frac{n_d}{t_d}\right)^{1/2}}{n - n_d}$$

where t and $t_d = 2\tau$ for a simple ratemeter circuit.

The sensitivity to terrestrial radiation is therefore dependent on the relative values of n and n_d, and the variation in the latter with time and position may determine the precision of measurement. It is therefore necessary to consider each contributing radiation in terms of the efficiency of its detection, its magnitude and intrinsic variation.

(3.1) Cosmic Radiation

Bombardment of nuclei in the earth's atmosphere by very energetic charged particles of external origin results in the emission of secondary radiation (mesons, photons and fast electrons) which can be detected near the earth's surface. Owing to the high energy of a large proportion of this radiation, the detection efficiency of any radiation detector approaches 100%.

Small variations in cosmic-ray intensity occur with time, but the major changes are with barometric altitude and to a lesser extent with geomagnetic latitude. These variations are usually negligible in a given survey area, and the cosmic contribution can be assumed constant in time and position.

(3.2) Atmospheric Radioactivity

Variations in the counting rate of γ-ray detectors have been observed which are in sympathy with the diurnal variations in the radon distribution in the atmosphere,[8] and have been attributed to this effect.[9] The decay of radon, the gaseous daughter of radium, gives rise to the solid decay products lead 214 and bismuth 214, which are also the principal γ-emitters of natural uranium. Thus the detection efficiency for atmospheric radiation can be considered approximately the same as for terrestrial radiation. The radon concentration near the ground changes markedly when a period of temperature inversion is followed by turbulent air conditions. It has been related to potential temperature gradient[10] and correlated inversely with wind-gust amplitude.[11] The greatest variations in this contribution are therefore likely to be experienced in regions where these meteorological conditions are possible and particularly if uranium mineral deposits are being mined in the same areas.

(3.3) Radiation Detector

The choice of detector is influenced by the relative contributions of the terrestrial and other radiations already discussed, since the sensitivity increases with n/n_d. The detection efficiency for cosmic radiation is similar for most types of detector, whereas for terrestrial and atmospheric γ-radiation, it varies in magnitude with detector type; therefore a highly sensitive γ-radiation detector will have a relatively small contribution from cosmic radiation. A sodium iodide [NaI(Tl)] scintillation counter can have a high γ-efficiency and has advantages over less efficient detectors. For example, at 1 m above chalk bedrock, cosmic radiation accounts for only a few per cent of the counting rate of a typical scintillation counter, whereas for a Geiger–Müller counter the contribution is 50%. This disadvantage associated with the Geiger–Müller counter was overcome in early equipments by using arrays of counters in coincidence and anticoincidence, but this resulted in a detector system of great complexity.

The scintillation counter has a small inherent counting rate arising from potassium 40 in the phosphor and glass envelope of the photomultiplier and from radium in other components[12]; this cannot readily be reduced. A greater contribution may be present if the counter is fitted in an aircraft having radioactive luminous instruments. By replacing these with inactive equivalents and installing the detector near the tail, this contribution can be minimized and is constant thereafter unless the aircraft becomes contaminated by airborne radioactive material.

(3.4) Radiation Components in a Scintillation Counter

Table 1 summarizes the contributions to the total counting rate in a typical NaI(Tl) scintillation counter situated 500 ft above ground. The datum level is large and can markedly change with time and position, owing to the large contribution from atmospheric activity. Variations with time are often greater than positional changes in a practical survey, but the amplitude of variation depends upon the location of the survey area. For example, in Rhodesia, which is highly mineralized, changes of the order 330 counts/sec have been observed,[9] compared with 100 counts/sec in Kenya[13] and 30 counts/sec in Britain.[2] It is therefore important to devise means of compensating for these time variations.

Provided that deposited contamination levels are insignificant, data of the natural activity in surface rocks can be obtained by subtracting the datum mean counting rate from the total mean counting rate. A measure of the deposited contamination is the primary object of an environmental survey, and this difference contains the contribution from natural activity. If the latter is large and variable with position, the contamination levels can be measured accurately only if a pre-incident survey has established this natural contribution. The accuracy of post-incident environmental survey is then determined by the accuracy achieved on the pre-incident survey. When the contributions to total counting rate from natural and deposited sources are the same order of magnitude, the precision of instruments and techniques must be very high.

(4) THE BASIC TECHNIQUE OF AERIAL RADIOMETRIC SURVEY

(4.1) Radiometric Instrumentation

Much of the development work on aerial survey has been concerned with devising instruments of high γ-sensitivity which are compact and reliable enough to operate satisfactorily in a light or medium aircraft. Research into aerial survey apparatus was started in Britain in 1946, and a multi-tube Geiger–Müller detector was tested over parts of Cornwall in 1950.[14] Simultaneous investigations were carried out in America[15] and in Canada, where the use of NaI(Tl) scintillation counters was first reported in 1952.[16] The development of large-volume phosphors firmly established the feasibility of aerial radiometric survey. Improvements in circuit technique and the use of cold-cathode valves improved stability and reliability,[17] and junction-transistor circuits[18–20] now allow compact highly sensitive instruments to be designed.

The instrument consists typically of a scintillation counter comprising a phosphor and photomultiplier, a ratemeter, power supplies and a recorder. A thallium-activated sodium-iodide crystal is the most suitable phosphor, although large crystals are

Table 1

SOURCES OF RADIATION AND TYPICAL CONTRIBUTIONS TO THE COUNTING RATE (SCINTILLATION DETECTOR)

Source of radiation	Contribution at 500 ft	Variations with time	Variations with position
	Counts/sec		
Terrestrial radioactivity			
Natural activity: deep sources of uranium and thorium series and potassium 40	200 (sands), 700 (granite)	Substantially constant	With surface geology
Deposited activity			
(*a*) From nuclear installations. Various fission products (iodine 131 most likely hazard)	290 from 1 $\mu c/m^2$ iodine 131. Negligible except in accident conditions	With decay rate of nuclides	With weather at time of release and with topography
(*b*) Fall-out from nuclear weapons. (See Section 6.2.) Zirconium 95, niobium 95, caesium 137	0–200 in U.K. between 1956 and 1959	With rate and magnitude of explosions, meteorological conditions and decay rates of nuclides. May also vary due to harvesting of crops, ploughing of land, etc.	With rainfall, land utilization and drainage characteristics of terrain
Datum radioactivity			
Cosmic radiation	7 at 500 ft above sea-level in U.K.	Substantially constant	With geomagnetic latitude (constant in latitudes >40°) and with barometric altitude (doubles in 10 000 ft)
Atmospheric activity (plus cosmic-induced soft component). Largely radon daughters	90–350	With wind gust amplitude and potential temperature gradient	With surface geology, mining activity and regional meteorology
Radioactivity of detector, aircraft instruments and structure	10–100	Constant for any aircraft unless it becomes contaminated	
Total datum counting-rate	Approximately 100–450	With atmospheric activity and from aircraft to aircraft	

expensive compared with the large blocks of organic phosphor which are sometimes used. Nevertheless, sodium-iodide total γ-detectors are less dependent upon temperature variations and exhibit a characteristic counting-rate/voltage plateau[17] which is not a feature of the organic phosphor. These advantages greatly simplify the problems of instrument design, and consequently NaI(Tl) phosphors have been used in the British instruments. Multiple scintillation detectors[20] avoid the disproportionate cost of larger crystals, give the required high sensitivity and have an advantage over single large detectors in reducing the chances of complete breakdown when airborne. The failure of one detector in a 3-detector system merely reduces the sensitivity of the instrument by 18%.

A sodium-iodide scintillation counter operating on plateau is sensitive to γ-radiation of energy greater than about 20 keV. The output pulses from the photomultiplier tube are converted to a constant width and amplitude by means of a trigger circuit, which feeds a ratemeter indicating mean counting rate. The degree of statistical fluctuation and the response time of the instrument can be selected by adjustment of the ratemeter circuit time-constant, which is normally variable between 0·5 and 5 sec.

(4.2) Aircraft Ground Clearance

The normal upper and lower limits of aircraft ground clearance are set by air attenuation and aircraft safety at 600 and 100 ft respectively. The maximum response from terrestrial activity can be obtained at the lower limit, but the flight lines must be closely spaced for adequate ground coverage. Thus low-level surveys (200 ft ground clearance and 500 ft spacing) are used for small-scale investigation of surface detail, but the large flight mileage per square mile makes the method too expensive for large-scale surveys, which are therefore usually carried out at 500 ft ground clearance and ¼-mile spacing.

In the United Kingdom the flying height is limited by statute and the normal minimum is 500 ft, since this is the closest approach permitted to any person, vessel, vehicle or structure.[21] This minimum height is increased to 1 500 ft over congested areas of cities, towns or settlements.

In practice, the aircraft cannot be maintained at an exact ground clearance, ±75 ft being typical deviations, and variations in detector counting rate occur which are not related to changes in terrestrial radioactivity; thus a correction method is normally necessary.

(4.3) Speed of Search

The greatest rate of change of γ-flux with distance is experienced when approaching or leaving a point source on the flight track, and for a given speed a short time-constant of the instrument will allow the recorded counting rate to follow the flux change faithfully. If the source is weak the recorded peak may then be indistinguishable from the statistical fluctuations on the counting rate preceding and following the flux change. Should the time-constant be made very long, the statistical fluctuations will be reduced but the recorded profile will be decreased in amplitude, broadened and delayed. The choice of survey speed and time-constant is therefore a compromise between these two conditions.

An analysis by Peirson and Franklin[14] gives, for each speed and flying height, a time-constant for the highest ratio of peak amplitude to statistical fluctuation. The recorded profile amplitude over a point source is then about 50% of the true flux change, while for an infinite-area source the recorded and true amplitudes are the same.[7] Thus maximum detectability of point sources is achieved, but absolute measurements of γ-flux (related, say, to iodine contamination levels) are made difficult, since the error varies with source dimensions. An alternative choice of time-constant can be made for each flying height by

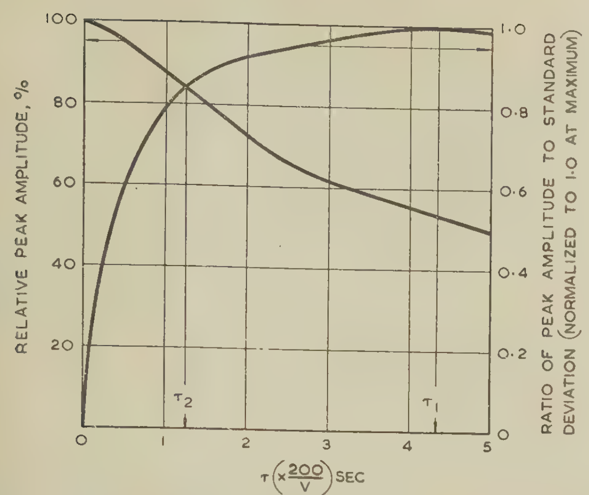


Fig. 3.—Diagram showing the choice of optimum $v\tau_2$ for 500 ft ground clearance.

compromising between the greatest signal to fluctuation and the highest ratio of recorded to true peak amplitude,[7] as shown in Fig. 3. Using these time-constants the maximum peak amplitude error is reduced to 17%, whilst the signal/fluctuation ratio is only worsened by the same percentage.

This condition is expressed for convenience in terms of the product of aircraft speed and time-constant, $v\tau_2$, for each flying height, since this product represents a displacement along the flight path. For example, at 500 and 200 ft ground clearance the product is 250 and 114 ft respectively. Thus for $v = 200$ ft/sec and $h = 500$ ft, τ_2 should be chosen at 1·25 sec. The recorded profile is substantially independent of the individual values of v and τ_2, but to minimize the statistical fluctuations, v should be small and yet compatible with practical requirements. A speed between 80 and 150 m.p.h. is normally acceptable, since it permits a choice from several aircraft, and allows accurate height-keeping and navigation at 200 to 500 ft ground clearance.

The increase of $v\tau_2$ with detector height demonstrates the point made earlier that the speed of aerial survey can be greater than that of ground survey without as serious a deterioration in point-source detectability as would be indicated by the γ-attenuation alone.

(4.4) Aircraft

An aircraft suitable for radiometric survey requires adequate power to maintain reasonably constant ground clearance, an ability to fly safely at 80–150 m.p.h., sufficient load-carrying capacity and a flight endurance of about 5 hours. Good forward visibility is essential for visual navigation, and the necessary instruments should be fitted to permit transit flights at night or a return to base in poor visibility.

Several aircraft meet these requirements, and although single-engined aircraft are cheaper to operate, twin or multi-engined types allow more flexibility in flight planning and are often inherently safer. Helicopters have advantages for detailed surveys at low altitude[22] and in mountainous areas. Their use is restricted in populated areas, and their lower speed, higher operating cost and limited flight endurance make fixed-wing aircraft preferable for most radiometric surveys.

(4.5) Position Plotting

An essential requirement in aerial survey is to correlate the radiometric record with the plan-position of the aircraft. An approximate indication of position can be obtained by direct visual identification of surface features, but the method is of restricted use since ambiguities often arise between landmarks. Where ground detail can be identified from maps or photomosaics, the flight track can be obtained by continuously photographing the terrain vertically beneath the aircraft. Automatic correlation of the film strip with the radiometric record provides a method which is inherently precise if the camera is gyrostabilized, otherwise slight errors can arise from random aircraft movements caused by air turbulence.[13] Over featureless country and when maps or photomosaics are not available, navigational systems such as Decca Navigator, Shoran or possibly Döppler can be used.

(4.6) Correction for Variations in Aircraft Ground Clearance

It is impossible to maintain the aircraft ground clearance at an exact value, and the greatest deviations occur in steeply undulating country. Changes in the recorded γ-flux occur, since the aircraft height above the nearest ground changes and because the solid angle subtended at the detector deviates from 2π, this effect being most marked during flights over ridges or along narrow valleys. The comparison of simultaneous indications from two detectors at different heights, e.g. one in the aircraft and one in a towed 'bird' at a lower height,[23] has been used to correct for these two effects. However, in all but the most severe topography, changes of solid angle have only a slight effect and can be neglected.

Full correction for changes in ground clearance requires a knowledge of all the source types and configurations encountered. Since this is not available, a partial correction can be derived from the measured law of γ-attenuation with height over flat, uniformly radioactive, surface rock. On geological surveys this gives good correction over large areas of radioactivity, and although it under-corrects the indications from small-area sources, it reduces the possibility of spurious anomalies arising from changes in ground clearance. The same correction can be applied on environmental surveys.[2] In this case the correction is accurate over uncontaminated areas, but in error where radio-elements have been deposited. The method, although imperfect, is, however, necessary, since the control of a post-incident survey relies on the comparison of its records over uncontaminated sections of flight line with the pre-incident records.

A continuous record of the aircraft ground clearance can be obtained from a radio-altimeter whose output is recorded on a multi-channel recorder with the radiometric information. The latter may then be corrected manually after the survey using a system of cursors. This is a slow and tedious process, and it is preferable to employ automatic methods which utilize the output of the radio-altimeter to modify the indication of the radiometric instrument according to an exponential function of height.[2, 24, 25] Correction over large-area sources to better than ±5% accuracy is possible depending on the calibration accuracy of the radio-altimeter.

(5) SURVEY METHODS

(5.1) Flight Planning for Geological Surveys

The choice of flight plan depends on the geological requirements. For economic reasons the survey should be completed quickly, but the chances of missing potentially valuable mineral occurrences should be minimized. Large areas of terrain are surveyed at a ground clearance of 500 ft with flight lines ¼ mile apart, and detailed studies of smaller areas are undertaken at 200 ft with a 500 ft line spacing. Variations on these techniques are sometimes possible in particular geological conditions, and, for example, a greater line separation can be used when the survey is for extensive disseminated deposits of uranium minerals.

In general, the flight lines are planned to form a parallel grid intersecting the regional geological strike at about 90°, but in mountainous terrain it may be necessary to follow the topographic contours.

(5.2) Flight Planning for Environmental Surveys

Surveys using road vehicles form the primary method of measuring ground contamination, since samples of vegetation must be collected for analysis. If weather conditions are favourable, aerial survey can give a rapid indication of contamination outside the range of the car-borne survey, and this may be particularly important in areas which are divided by stretches of water or lack a good road network. Even if the aerial survey cannot be undertaken within 24 hours of an accident, information gained some time later can still be of value.[26]

A method of applying aerial-survey techniques to environmental survey has recently been devised and assessed.[2] The flight lines formed a pattern of three approximately circular lines, at 15, 30 and 45 miles radius, and four diametric lines, all centred on the atomic-energy station (Fig. 5), so that they could be flown in four sorties of less than 5 hours' duration each. Where practicable, flight lines were sited along hill sides, on which deposition might occur. The method relies on comparison of the results of a survey flown after a release of activity with those of a pre-incident survey occupying the same flight lines. The lines were therefore planned for ease of visual navigation, following landmarks such as roads, railways, power lines and avoiding steep undulations in the terrain.

(5.3) Datum Determination

Ideally, a continuous measurement of the terrestrial radiation intensity should be made as the survey proceeds, and this would be possible if the datum component could be isolated. No means of achieving this has been found and therefore a method of measuring datum changes with time has been devised.[9] This employs one or more flight lines which are accurately re-occupied at frequent intervals during the survey, and the variation with time of the radiation intensity over each line is plotted. If the lines are sited over flat terrain and arranged to follow well-defined ground features, they can be flown with precision in both plan-position and ground clearance. Since the terrestrial radiation contribution can be assumed constant for all such control-line flights, observed variations can be attributed to datum variations if the radiometric instrument is sufficiently stable. Provided that a control line is characteristic of the survey area, similar variations in datum can be assumed to occur over the area and appropriate corrections made to the counting-rate records.

The precise control system adopted for any area depends upon the survey requirements. For example, in uranium prospecting a simple system can be used, since the search is concerned with detecting localized anomalous activity, and the broad radiation levels over the country rocks are not required to be accurately known. On the other hand, more precise control must be exercised for geological mapping and for surveying low concentrations of ground contamination.

Control lines should be sited to minimize unproductive flying, and the simplest system uses a single control line sited in the same geological and topographic setting as the survey area.[9, 13] The mean counting rate over this line is measured before and after each sortie, and the datum value is assumed to vary linearly with time between the two measurements. Thus the correction applied to each survey line record is obtained from the difference between the lowest recorded control-line value and the interpolated value for the time at which the survey line was flown.[9] This system gives no direct information of the variations in datum either during each sortie or with position over the survey area. For a large area the latter defect can be mitigated by the use of more than one control line.

When the region to be surveyed is long and narrow, a control line can be sited centrally along the length of the area[19] and flown in its entirety at the beginning of the survey. That section of control-line intersected by the survey lines on any one sortie is re-occupied at the beginning and end of the sortie, and a comparison of the mean counting rates with those originally obtained over the same section provides the necessary corrections. This system makes allowance for regional datum variations, and some account can be taken of datum changes occurring during a sortie by inspection of the values obtained where each survey line intersects the control line.

An improved system for geological survey is to site control lines at the edges of the survey area. To aid navigation, survey lines are often divided into groups, so that all lines in each group are flown in the same direction. Transit flights between lines can then be employed for control purposes (Fig. 4), thereby

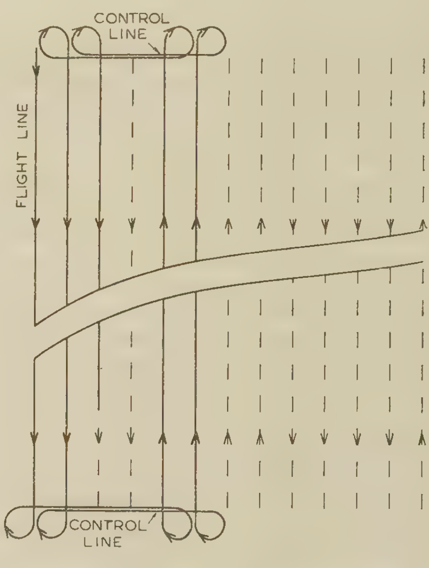


Fig. 4.—A control-line system for aerial geological surveys.

utilizing what would otherwise be unproductive flying. The area should have recognizable features, such as roads, railways and rivers, as boundaries, so that the control lines can be readily re-occupied. Good control is established at the beginning and end of each survey line, and the additional advantage ensues that extensions to the survey area can employ one of the original control lines.

A similar approach can be made to the control of environmental surveys. Since the flight pattern is basically circular, geological-type controls cannot easily be employed and a system has been based[2] on comparison of records obtained during re-occupation of 'common' sections of flight line on successive sorties (Fig. 5). These are supplemented by noting the counting rates at points where lines intersected, and consequently the intersection points were sited where possible over flat, homogeneously radioactive, surface rock.

These systems of control take reasonable account of datum changes and give self-consistent results in a given survey area, but do not provide an absolute measure of the datum level, which may vary from survey to survey. Surveys have been related by measuring the total counting rate at places in or near the survey area where the terrestrial contribution is reduced to

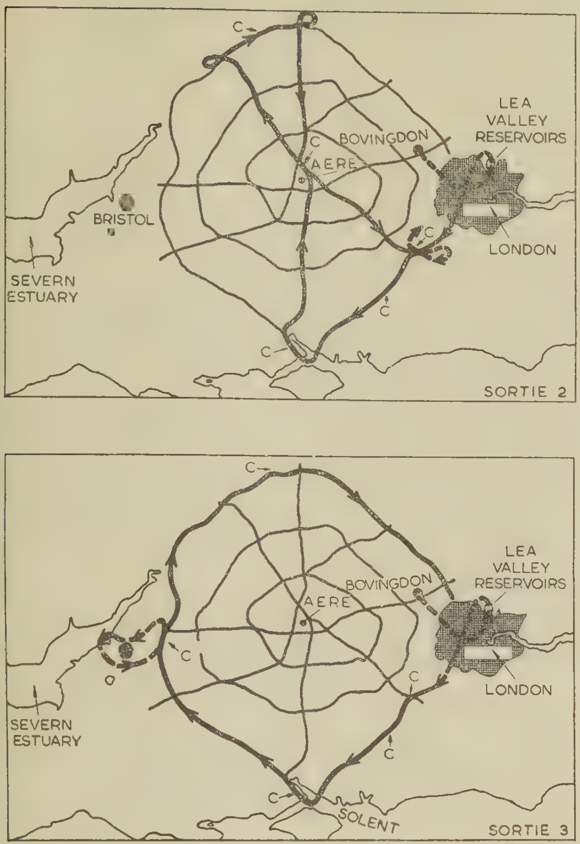


Fig. 5.—A typical environmental-survey flight plan.

The plan was used for an experimental survey around A.E.R.E. Harwell and consists of a system of approximately circular flight lines centred on the establishment. The aircraft track on two sorties is shown by a thickened line marked with arrows showing the direction of flight. Each sortie's track occupies some sections of line flown during that sortie and on other sorties. The records for these common lines (marked C), together with those for intersection points, were used for assessing variations in datum level. The residual radiation level was established by flights over the stretches of water shown.

negligible proportions, e.g. above an extensive stretch of water. This 'residual' counting rate, comprising the inherent radioactivity of the instrument and aircraft, cosmic radiation, and small, substantially constant, but unknown, contributions from atmospheric activity, is subtracted from the survey-line counting rates after control-line corrections have been applied, to give a more nearly representative value for the true terrestrial radiation levels. This method is obviously imperfect, since the atmospheric activity over water is likely to be different from the minimum recorded activity over the control line,[2] but practical evidence so far suggests that it is not seriously in error. Where no suitable stretches of water are available, the lowest reading observed at 2000 ft above ground can be taken as the 'residual' level with somewhat greater error.[13]

(5.4) Presentation of Results

The presentation of the very large amount of data compiled during a survey should give the maximum information in the most compact form. The starting-point is a map showing the flight tracks and a radiometric record corrected for ground-clearance variations and marked with a datum line. Where the flight lines are straight and widely separated, which is rarely the case, the radiometric profiles, correctly scaled, can be placed along them. The more usual method employs a counting-rate scale divided above datum into intervals based on counting statistics. The interval increases with total counting rate, n, and is three times the standard deviation of n at its lower limit.[2] The points where the record profile crosses from one interval to another are transferred to the mapped flight-lines. 'Iso-rad' contours can then be made up by joining points of equal radiation intensity, but uncertainties in the interpolation between flight lines makes this system difficult to interpret; moreover, significant anomalies may be missed or insignificant ones recorded. These difficulties can be overcome by separate identification and marking of the anomalies, and by marking the general radiation levels along the flight lines. Appropriate colours are used for the counting-rate intervals, and each anomaly is indicated by a circle, coloured according to the peak amplitude above the adjacent background, together with a number code giving its width at half-height in terms of camera frames.[9] An example of this system (without the colours) is shown in Fig. 6.

A similar method has been adopted for presenting environmental survey results[2] (Fig. 7). In this case the flight track is marked as a broad line of width approximately equal to the effective ground cover, divided into coloured sections as before, and numbered with the appropriate interval. Since broad changes of level are now of primary importance, local but significant peaks and troughs in the record are marked less conspicuously as red and black lateral lines respectively.

(6) INTERPRETATION OF DATA

(6.1) Anomalies

Since the radiation intensity above a radioactive ore body is so dependent upon its thickness and depth below the surface, the amplitude of an anomaly can rarely be related directly to ore-grade. Consequently, in the search for uranium, any statistically significant peak on the record may be potentially important, and the main value of aerial survey is therefore to direct ground survey teams to places which are worthy of detailed examination.

Provided that suitable time-constants and aircraft speeds are chosen, the configuration of the source of radiation can often be deduced. For example, a sharp peak appearing only on one flight-line may indicate the presence of a small-area source, whereas similar anomalies on a number of adjacent flight-lines may be due to an elongated source across the direction of traverse. Similarly, a broad peak on one line implies a linear source parallel to the flight direction, and similar profiles on adjacent lines imply an extensive area of anomalous radioactivity.

A 'total-γ' detector of the type described is incapable of distinguishing between anomalies due to uranium minerals, thorium or radio-potassium. A scintillation counter can, of course, resolve differences in γ-energy spectra by pulse-height analysis, and in recent years airborne apparatus using this principle has been used,[27, 28] but no reports of the practical value of the method are at present available.

(6.2) Ambiguities in Data Interpretation

The assumption made earlier that the radiation from uncontaminated ground is constant in time is not strictly correct in practice. Variations can occur due to changes in γ-absorption caused by the variable water content of the soil[29] or the presence of snow. The former effect is most marked over low-lying land which may become saturated or flooded, and, in fact, such localities can readily be identified by local troughs in the radiometric record.[2]

In recent years a more serious ambiguity has arisen from fall-out from nuclear-weapon trials, although, with the cessation of tests, the problem has become less significant.[2] Part of the fall-out is retained by vegetation and the rest accumulates in the

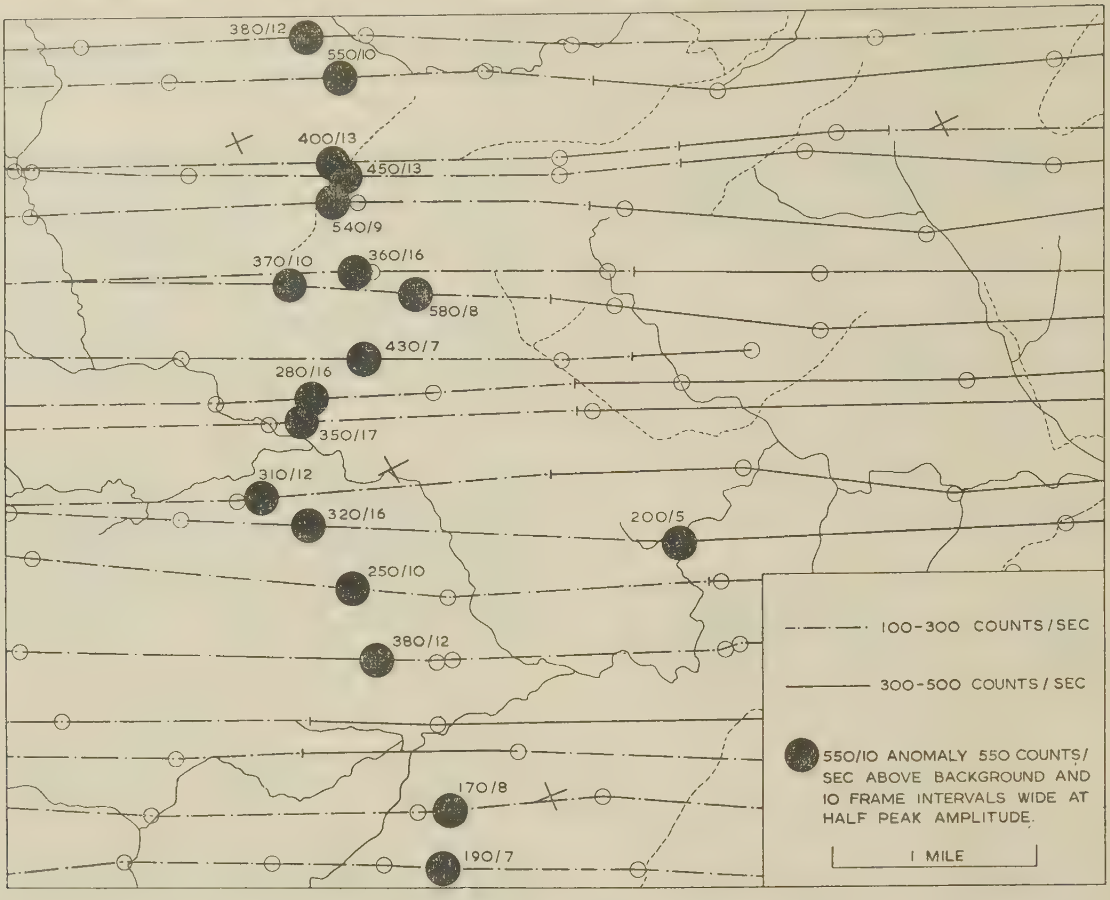


Fig. 6.—Presentation of aerial geological survey data.

top few centimetres of soil, the surface concentration varying in time with the frequency and magnitude of test explosions, meteorological conditions and the rate of decay of the constituent radio-elements. Positional variations also occur, owing to differences in rainfall, land utilization and the drainage characteristics of the terrain. The main sources of radiation have been zirconium95 and niobium95, which reached peak concentrations in April, 1959, and at that time their γ-contribution was comparable with that arising from the natural radioactivity of sedimentary rocks.[30] High readings, many times the local background, were also observed at places where the material had been concentrated, e.g. near drainage channels and hayricks.

Fall-out can most seriously affect the precision of measuring broad changes of terrestrial radiation levels, as in pre-incident environmental surveys, and total-γ techniques do not permit any method of discrimination. It is therefore necessary to repeat surveys if a marked change of fall-out concentration is indicated by monitors on the ground. A search for anomalies, as in uranium prospecting, is not so seriously affected, although, as mentioned above, local ambiguities can arise.

(7) CONCLUSIONS

Aerial radiometric survey has been established as an effective and economic method of uranium prospecting in areas which are geologically hospitable to uranium minerals.[19] It can be used for mapping regional geology, and this application has been reported to be of value in locating oilfields by virtue of an increased γ-radiation intensity around the periphery of the oil-bearing structure.[31] Absolute γ-intensities, as measured over large-area sources on a closely controlled aerial survey, can be related within $\pm 15\%$ to those observed on the ground.[2]

An aerial survey of ground contamination following a release of γ-emitters from an atomic-energy site can be a valuable supplement to ground surveys, and may indeed be the only rapid and economic method for the survey of areas more than 15 miles from a site where the road network is restricted by stretches of water or moorland. Given adequate control, the precision of the method is such that a surface concentration equivalent to less than 1 $\mu c/m^2$ of radio-iodine (iodine131) can be detected.[2]

Present knowledge suggests that the sensitivity of radiometric instruments developed at the Atomic Energy Research Establishment[20] is adequate for geological and environmental surveys. Although greater sensitivity could be achieved by virtue of the improved counting statistics resulting from the use of larger phosphors, inherent imperfections in the methods of altitude and datum correction indicate that correspondingly more accurate determinations of terrestrial radioactivity would not be achieved, and the resulting increase in physical dimensions and capital cost of the equipment is not considered warranted.

The stability of modern instruments, using inorganic phosphors operating on plateau, is better than $\pm 2\%$ over long periods, and the use of transistor circuits has made their reliability extremely high.[2–13]

Fundamental improvements in aerial radiometric survey methods are therefore dependent upon finding improved means

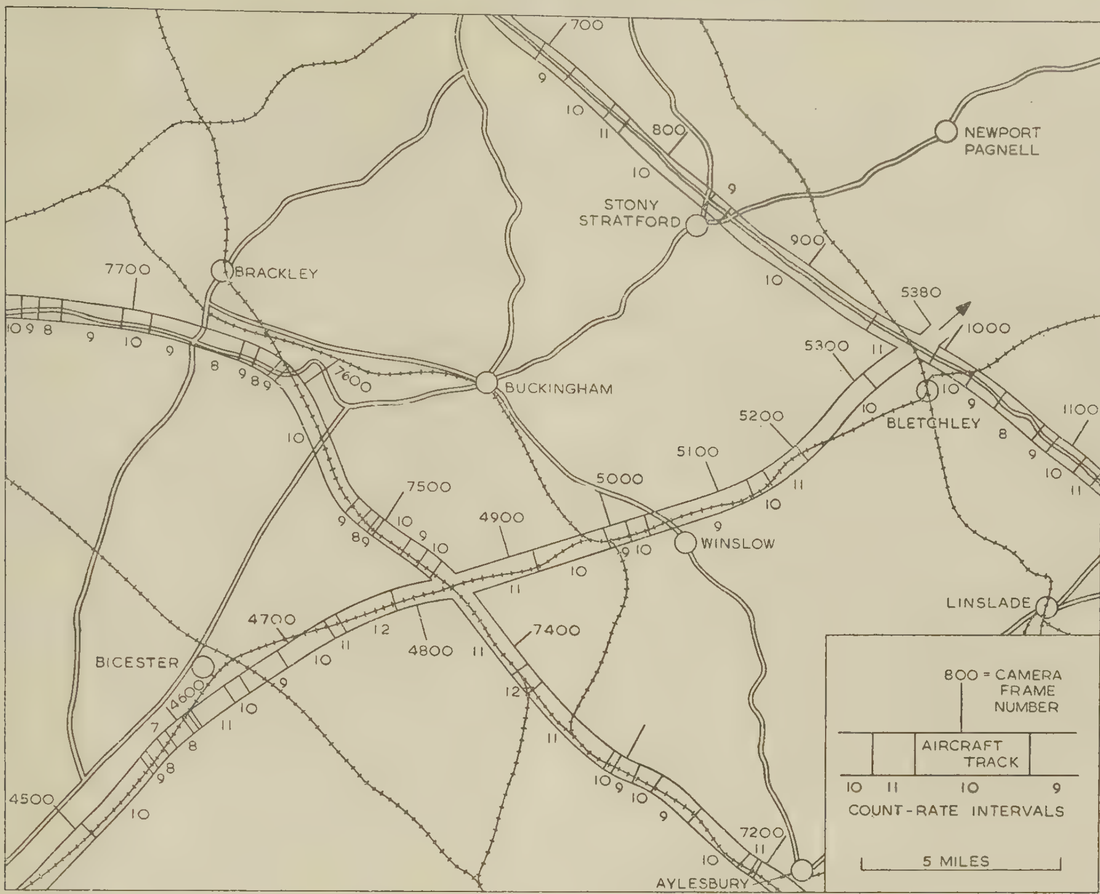


Fig. 7.—Presentation of aerial environmental survey data.

of discriminating against unwanted radiation, more precise flying, or of correcting for aircraft ground-clearance variations. γ-spectrometry could, in principle, be used on environmental surveys for identifying the predominant γ-emitters in the main areas of contamination. Aerial survey would be greatly improved if more rapid and cheaper data reduction methods were available, and this may rest on improved systems of correlating the radiometric record with aircraft plan-position.

(8) ACKNOWLEDGMENTS

The authors wish to record their appreciation of the advice given by Mr. E. H. Cooke-Yarborough in preparation of the paper, the supporting work done by Messrs. R. S. Cambray and J. Pickup, all members of Electronics Division, A.E.R.E., Harwell, and also the co-operation of Hunting Surveys Limited in execution of the flying programmes to which reference is made in the text.

(9) REFERENCES

(1) Farmer, F. R., and Fletcher, P. T.: 'Siting in Relation to Normal Reactor Operation and Accident Conditions', Symposium on Safety and Location of Nuclear Plants, Rome, 1959.

(2) Williams, D., and Cambray, R. S.: 'Environmental Survey from the Air', A.E.R.E. Report R–2954, 1960.

(3) Davis, F. J., and Reinhardt, P. W.: 'Instrumentation in Aircraft for Radiation Measurements', *Nuclear Science and Engineering*, 1957, **2**, p. 713.

(4) Spencer, L. V., and Fano, U.: 'Penetration and Diffusion of X-Rays: Calculation of Spatial Distribution by Polynomial Expansion', *Journal of Research of the National Bureau of Standards*, 1951, **46**, p. 446.

(5) Goldstein, H., and Wilkins, J. E.: 'Calculations of the Penetration of Gamma Rays', U.S. Atomic Energy Commission, Report NYO–3075, 1954.

(6) Sakakura, A. Y.: 'Scattered Gamma Rays from Thick Uranium Sources', U.S. Geological Survey Bulletin 1052–A, 1957.

(7) Williams, D., and Cambray, R. S.: 'The Selection of Parameters for Aerial Radiometric Survey', A.E.R.E. Report R.3433, 1961.

(8) Becker, F.: 'Measurement of the Emanation Content of Air at the Meteorological Institute at Frankfurt on the Main', *Gerlands Beiträge zur Geophysik*, 1934, **42**, p. 365.

(9) Cambray, R. S., Williams, D., and Bisby, H.: 'An Aerial Radiometric Survey of the Urungwe District, Southern Rhodesia', A.E.R.E. Report EL/R–2571, 1958.

(10) Gale, H. J., and Peaple, L. H. J.: 'Measurements of the Near-Ground Radon Concentration on the A.E.R.E. Airfield', A.E.R.E. Report HP/R–2381, 1958.

(11) Wilkening, H.: 'Daily and Annual Courses of Natural Atmospheric Radioactivity', *Journal of Geophysical Research*, 1959, **64**, p. 5.

(12) Miller, C. E., Marinelli, L. D., Rowland, R. E., and Rose, J. E.: 'An Analysis of the Background Radiation Detected by NaI Crystals', *Transactions of the Institute of Radio Engineers*, 1956, NS–4, p. 90.

(13) Cambray, R. S., and Hill, W. G.: 'An Aerial Radiometric Survey of Parts of Kenya and Tanganyika', A.E.R.E. Report R–3224, 1960.

(14) Peirson, D. H., and Franklin, E.: 'Aerial Prospecting for Radioactive Minerals', *British Journal of Applied Physics*, 1951, 2, p. 281.

(15) Stead, F. W.: 'Instruments and Techniques for Measuring Radioactivity in the Field', First International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1955, Paper No. 511.

(16) Steljes, J. F., Cowper, G., and Carmichael, H.: 'Aerial Prospecting for Radioactive Materials', National Research Council of Canada Report No. CRR–495, 1952.

(17) Peirson, D. H., and Pickup, J.: 'A Scintillation Counter for Radioactivity Prospecting', *Journal of the British Institution of Radio Engineers*, January, 1954, **14**, p. 25.

(18) Pickup, J., and Cosgrove, M. E.: 'A Scintillation Counter for Airborne Radiometric Surveying', A.E.R.E. Report EL/R.1972, 1956.

(19) Bowie, S. H. U., Miller, J. M., Pickup, J., and Williams, D.: 'Airborne Radiometric Survey of Cornwall', Second International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958, Paper No. 43.

(20) Pickup, J.: 'Instrumentation for Aerial Radiometric Surveys', A.E.R.E. Report R–2953, 1961.

(21) Civil Aviation, Air Navigation Order 1954, No. 829, Schedule II. Sections IV, V, p. 64, and Amendment Order No. 82, 1956.

(22) Ziehr, H. Von.: 'Uranprospektion mit dem Helicopter', *Die Atom Wirtschaft*, 1960, **5**, p. 223.

(23) Frank, E. J.: 'An Airborne Computer-Controlled Detector for Radioactive Ores', *Journal of the British Institution of Radio Engineers*, November, 1956, **16**, p. 11.

(24) Cassidy, M. E., Graveson, R. T., and Levine, H. D.: 'HASL Aerial Survey System', U.S. Atomic Energy Commission, Report NYO–2071, 1957.

(25) Dmitriev, A. V., and Ionov, V. A.: 'L'Automatisation de la correction (d'altitude) au cours des leves aeriens du rayonnement gamma', *Geophysique*, 1958, **4**, p. 31.

(26) Williams, D., Cambray, R. S., and Maskell, S. C.: 'An Airborne Radiometric Survey of the Windscale Area, October 19th–22nd, 1957', A.E.R.E. Report R–2890, 1959.

(27) Berbezier, J., Blangy, B., Guitton, J., and Lallemant, C.: 'Méthodes de prospection autoportée et aeroportée: la technique de la détection des rayonnements: les perspectives offertes par la discrimination des énergies', Second International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958, Paper No. 1249.

(28) Matveev, V., and Sokolov, A. D.: 'An Aircraft Radiometer Analyzer', *Atomnaya Energya*, 1960, **8**, p. 70.

(29) Macfadyen, D. A., and Guedes, S. V.: 'Air Survey Applied to the Search for Radioactive Minerals in Brazil', First International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1955, Paper No. 132.

(30) Peirson, D. H., and Salmon, L.: 'Gamma Radiation from Deposited Fall-Out', *Nature*, 1959, **184**, p. 4700.

(31) Lundberg, H., and Isford, G.: 'Oil Prospecting with the Radioactive Method', *World Petroleum*, July, 1953, **29**, p. 7.

621.374.32 : 621.317.79 : 621.039

The Institution of Electrical Engineers
Paper No. 3535 M
July 1961
©

THE CONTINUOUS MEASUREMENT OF THERMAL-NEUTRON FLUX INTENSITY IN HIGH-POWER NUCLEAR REACTORS

By W. R. LOOSEMORE, B.Sc., and J. A. DENNIS, B.Sc., A.Inst.P.

(*The paper was first received* 10*th November, and in revised form* 28*th December*, 1960.)

SUMMARY

The problem of measuring continuously the thermal-neutron flux intensity in high-power nuclear reactors at temperatures up to 500° C is considered, and it is concluded that the mean-current ionization chamber provides the best solution.

The development of a suitable chamber is described, which is made from stainless steel and has a gas filling of xenon. When the electrodes are coated with uranium oxide (U_3O_8) containing 0·48 mg of uranium 235, the current sensitivity is 7×10^{-17} A/n/cm²/sec, and at a neutron flux intensity of 10^{13} n/cm²/sec the equilibrium residual current measured in a graphite-moderated reactor is 0·6% of the total output. Life tests have been carried out successfully to a total neutron dose of 3×10^{20} n/cm².

At a flux intensity of 10^{14} n/cm²/sec a chamber with uncoated titanium electrodes has the same fraction of residual current as one containing uranium 235 and is preferred as it is not subject to depletion effects.

LIST OF SYMBOLS

$V_{0\cdot 9}$, $V_{1\cdot 1}$ = Potentials at which chamber currents are respectively 0·9 and 1·1 of saturation current, volts.
I_s = Chamber saturation current, amp.
p = Gas pressure, mm Hg.
d = Electrode spacing, cm.
α = Number of electrons produced in the path of a single electron travelling a distance of 1 cm in the direction of the electric field.
E/p = Reduced electric field, volts/cm per mm Hg.
W_s, W_g = Energy absorptions per unit mass of solid and gas respectively, eV/g.
S_s, S_g = Mass stopping powers of solid and gas, respectively, for β-radiation, eV-cm²/g.
N_g = Number of ion pairs formed per unit mass of gas, pairs/g.
w = Mean energy required to produce one ion pair, eV.
σ_{a5}, σ_{a8}, σ_{a9} = Effective neutron absorption cross-sections per atom, for uranium 235, uranium 238 and plutonium 239, respectively, barns.
σ_{f5}, σ_{f9} = Effective fission cross-sections per atom, for uranium 235 and plutonium 239, respectively, barns.
$\bar{\sigma}_f$ = Effective fission cross-section per atom of mixture at time t, barns.
R = Initial ratio (number of atoms of uranium 238)/(number of atoms of uranium 235).
ϕ = Thermal-neutron flux intensity according to the Westcott convention, n/cm²/sec.

(1) INTRODUCTION

Nuclear reactors at present under construction as part of the nuclear-power programme in this country will operate with a maximum core temperature of 300–400° C, and in future designs this may be increased to 500° C; the thermal-neutron flux intensity in such a reactor will be about 10^{13} n/cm²/sec and may eventually approach 10^{14} n/cm²/sec. Normally the reactor power output will be controlled at the desired level by measurement of the thermal-neutron flux intensity in a thermal column outside the core, using standard boron-lined mean-current ionization chambers.[1] However, it is also desirable to obtain information about the thermal-neutron flux intensity inside the reactor core, and two types of measurement may be distinguished in this context. The neutron flux intensity through the core may be scanned at regular intervals as a check that the desired distribution is maintained at start-up and subsequently throughout the life of the reactor fuel charge; this type of measurement requires a neutron detector which can be moved rapidly and easily through the core. Additionally, it may be desirable to have detectors permanently sited at a number of points in the core to monitor continuously the spatial distribution of flux intensity; a continuous monitoring system of this type might be used in the control of spatial instabilities in the operating state of the reactor.

Any instrument for 'in-pile' use will be required to function reliably in a neutron flux intensity of at least 10^{13} n/cm²/sec at temperatures up to 500° C for periods of not less than one year. If it is to detect instabilities it must detect a step change in flux intensity of 1%, over a range of 10 : 1, in less than 30 sec, and its sensitivity should not vary by more than 5% during its lifetime in the reactor. It must be robust, particularly if it is to be used for flux scanning, so that it may be loaded into the reactor by the standard fuel-element charging apparatus without reducing power. It should have a diameter of less than 2 in and it must be able to deliver a signal over a distance of up to 50 ft from the nearest accessible point through a connection which must be at least semi-flexible.

Consideration of the various devices which measure neutron flux intensity lead to the conclusion that some form of mean-current fission ionization chamber is most likely to fulfil the specification, but the possible alternatives are mentioned briefly. Any instrument which relies on the $^{10}B(n, \alpha)^7Li$ reaction to detect neutrons will be unacceptable in this application because of rapid depletion due to the high absorption cross-section of boron 10. In six weeks at a neutron flux of 10^{13} n/cm²/sec the sensitivity of such a device would decrease by about 20%. The fission reaction in uranium 235 has a lower cross-section, and the same reduction in sensitivity of a detector based on this reaction would occur in 1 year.

Pulse-type fission ionization chambers detect the presence of individual neutrons from the kinetic energy of the fission fragments (maximum 80 MeV for each fragment), but in a flux intensity of 10^{13} n/cm²/sec it is almost impossible to apply an amount of fissile material small enough to allow individual fission events in the detector to be separated in time: the fission rate in 1 μg of uranium 235 would be about $1\cdot 5 \times 10^7$ fissions/sec. Moreover, the high level of γ-radiation intensity will deposit large amounts of energy in the chamber, and fluctuations in this rate of deposition within the resolution time of the detecting system will produce signals which will be indistinguishable from those due to fission events. A further general disadvantage of

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Loosemore and Mr. Dennis are at the U.K.A.E.A. Atomic Energy Research Establishment, Harwell.

pulse detecting systems for high-temperature in-pile measurement lies in the high electrical capacitance of the connecting cables. The normal maximum amplifier input capacitance allowable with the best present-day techniques is about 500 pF. and cables designed for high-temperature operation commonly have a capacitance per unit length of 60 pF/ft, so that the maximum length of cable which can be used is about 10 ft; this is too short to be useful in a large reactor. The development of miniature high-temperature pulse-transformers[2] may overcome this cable-length difficulty in the future, but the problem of response to the high γ fluxes would remain.

Neutron-sensitive thermopiles[3, 4] do not require connecting cables of low capacitance or high electrical resistance, but they are difficult to make with an output which is linear over the required range. Their response time is often long, and although this may be reduced by suitable design, the resulting instrument is fragile.

It has been suggested that in-pile neutron measurements could be made by passing a gas stream over a small uranium sample in the reactor and measuring the quantity of gaseous radioactive fission products swept out by the carrier gas.[5] The rate of evolution of the fission products is proportional to the neutron flux intensity, but it is also sensitive to the temperature of the uranium sample.[6] It would be difficult to achieve a time-constant of less than a few seconds with this technique, and it would not be feasible to adapt it for flux scanning. As a variation of this method it has been proposed that argon be passed through the reactor and the neutron-induced radioactivity measured. The induced activity is proportional to the integrated flux intensity along the gas channel, but the measurement can be localized to some extent by transferring the gas through small-bore tubing and employing a relatively large hold-up volume at the point where the measurement is required. The same objections apply to this method as to the gaseous fission-product technique.

Activated-wire methods have been described[7] which are suitable for flux scanning, but the long time interval necessary before the induced activity can be assayed makes them unsuitable for the detection of flux instabilities.

The mean-current ionization chamber has several advantages over the neutron detectors mentioned above. The response time is usually governed by the current-measuring instrument to which it is connected, so that it may be only a few milliseconds. The chamber will operate with long connecting cables, and since it can be made very robust, it is suitable for permanent in-pile installation or for flux scanning. An ionization chamber which does not contain fissile material will respond to the prompt γ-radiations from (n, γ) reactions in the reactor core and in the materials of the chamber, and will therefore be neutron-sensitive. There will also be a contribution to the chamber output due to ionization produced by γ-radiation from fission product activity of the reactor core and from β- and γ-induced activity in the materials of the chamber. This residual respons will limit the range of neutron flux intensity over which th chamber can be used.

When the electrodes are coated with a thin layer of uranium 23 the neutron response is enhanced by the ionization produced b fission fragments released from fission reactions occurring i the coating. This type of chamber will have an output which i proportional to neutron flux intensity over a wider range o intensities than will a chamber with uncoated electrodes, bu the coating may be subject to radiation damage under the pro longed exposures which are envisaged. Both types of chambe were investigated and are discussed in detail below; it is con cluded that the chamber containing uranium 235 offers th most attractive solution to the problem for neutron flux inten sities up to 10^{13} n/cm^2/sec. At higher flux intensities it will b shown that a chamber with plain electrodes may becom increasingly attractive.

(2) ELECTRODE GEOMETRY AND GAS FILLING

The general form of the relationship between the curren collected from an ionization chamber and the potential differenc applied between its electrodes is known as the saturation charac teristic. As the applied potential is increased from zero th collected current increases rapidly at first, and then more slowl until a region is reached where it is nearly or completely inde pendent of the applied potential. This region is known as th 'plateau' of the characteristic and the current is the 'saturatio current', I_s. At higher potentials the current increases as th electrons gain sufficient energy from the electric field betwee the electrodes to make ionizing collisions with gas molecule and finally an arc discharge occurs. The plateau may be define conveniently by the values of the potentials $V_{0 \cdot 9}$ and $V_{1 \cdot 1}$ a which the current is 90% and 110% of the saturation value. A a general rule the maximum saturation current which can b drawn with complete reliability from a chamber is reached whe $V_{0 \cdot 9} = \frac{1}{3} V_{1 \cdot 1}$; an operating voltage of $V_0 = 2V_{0 \cdot 9}$ then give virtually 100% efficiency for the collection of ionization betwee the electrodes.

A considerable amount of theoretical and experimental wor has been published on the characteristics of ionization chambe in which the ionization is due to electrons or α-particles,[1] but no work has been reported on chambers in which th ionization is due to fission fragments. In order to establi data on which to base a final design, experiments were carri out on three chambers containing uranium 235 to study th effects of different gases, pressures and electrode spacings o the saturation characteristics. A cylindrical electrode structu was chosen as being the most convenient practical arrangemen

The inner-electrode diameter in each chamber was 2·0 c and the diameters of the outer electrodes were chosen to gi electrode spacings of 0·5, 1·0 and 2·0 mm. The chambers h the same general form as that shown in Fig. 1, but the spec

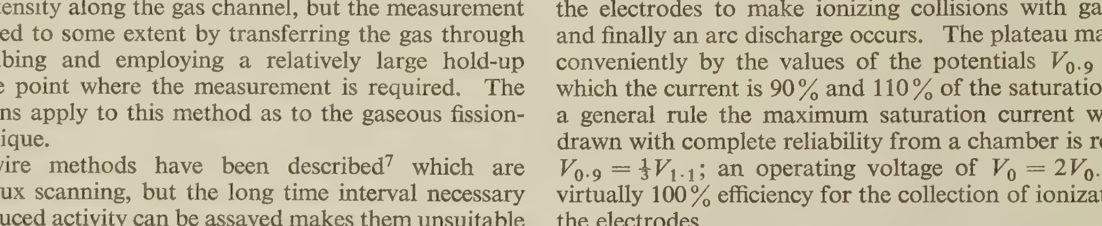


Fig. 1.—Single-containment guard-ringed ionization chamber.

A. Collector electrode. B. Guard-ring electrodes. C. H.T. electrode. D. Alumina plugs. E. Stainless-steel braid. F. Magnesia-insulated cables. G. Metal-alumina seals. H. Alumina insulators. I. Filling tube.

able-sealing arrangements were omitted. Guard-ring electrodes vere incorporated to maintain a uniform electric field along the ength of the collecting space and to minimize leakage currents. The outer surface of each collector electrode was coated with ranium oxide containing about 2·5 mg of uranium 235 deposited ver an area of 25 cm².

The three chambers were irradiated in turn in a thermal-ieutron flux intensity of up to 5×10^{10} n/cm²/sec in the Lido eactor, and were filled with neon, nitrogen or xenon at pres-ures between 140 and 2 280 mm Hg. The gases were chosen ecause they were known to have negligible electron-attachment oefficients and also because the neutron activation induced luring the experiment would not result in any great radiation azard when pumping them out.

The results with a 1 mm electrode spacing and a xenon filling re shown in Fig. 2, in which $V_{0\cdot9}$ is plotted as a function of the aturation current density per square centimetre of collector lectrode area, for different gas pressures and reactor powers; he value of $V_{1\cdot1}$ for each pressure is also included. Similar urves were obtained with electrode spacings of 0·5 and 2·0 mm ind with neon and nitrogen.[9] The reactor powers are nominal, ut the lines connecting points of constant power level are hown because they give information on the variation of chamber urrent with gas pressure for a constant neutron flux intensity.

It has not been found possible to develop a quantitative heoretical treatment describing the variation of $V_{0\cdot9}$ as a unction of I_s, p and d, but the general shape of the $V_{0\cdot9}/I_s$ urves may be explained. At low current densities the ionization roduced in the chamber gas by each recoiling fission fragment s independent of that due to other fragments, and columnar ecombination between ions in any one track is the most mportant cause of the removal of ions. In this region $V_{0\cdot9}$ will be substantially independent of I_s, although diffusion of ons out of the collecting volume may be expected to account or a slight increase in $V_{0\cdot9}$ as I_s increases. At higher current lensities recombination will occur between ions in different racks, and in this region there will be an increasing dependence of $V_{0\cdot9}$ on I_s. The effect will be more marked at the lower gas ressures, since the range of secondary ionizing particles in each track will be correspondingly increased. These features are borne out by the results shown in Figs. 2 and 3.

A value for $V_{1\cdot1}$ can be deduced from

$$\frac{I}{I_s} = \frac{(\varepsilon^{\alpha d} - 1)}{\alpha d} \qquad . \quad . \quad . \quad . \quad . \quad (1)$$

This expression represents the growth of the current I between plane parallel electrodes when the gas between them is uniformly ionized. Strictly, the equation holds only for very low ionization rates, when there is no space-charge distortion of the field. At $V_{1\cdot1}$, when $I/I_s = 1\cdot1$, $\alpha d = 0\cdot187$, so that

$$\frac{\alpha}{p} = \frac{0\cdot187}{dp} \qquad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Meek and Craggs[10] tabulate α/p as a function of the reduced electric field, E/p or V/dp, for neon, nitrogen and xenon, and by using their data in conjunction with eqn. (2) the appropriate value of $V_{1\cdot1}/dp$ can be calculated.

It can be shown that the error introduced by applying eqn. (1) to a cylindrical geometry is not more than a few per cent for the electrode dimensions which were used in the experiments, and since it is observed that $V_{1\cdot1}$ is very nearly independent of the saturation current density up to at least 4×10^{-5} A/cm², it may be inferred that the correction for space-charge distortion s also small. Very good agreement is obtained between the values of $V_{1\cdot1}$ deduced from the equation and the experimental

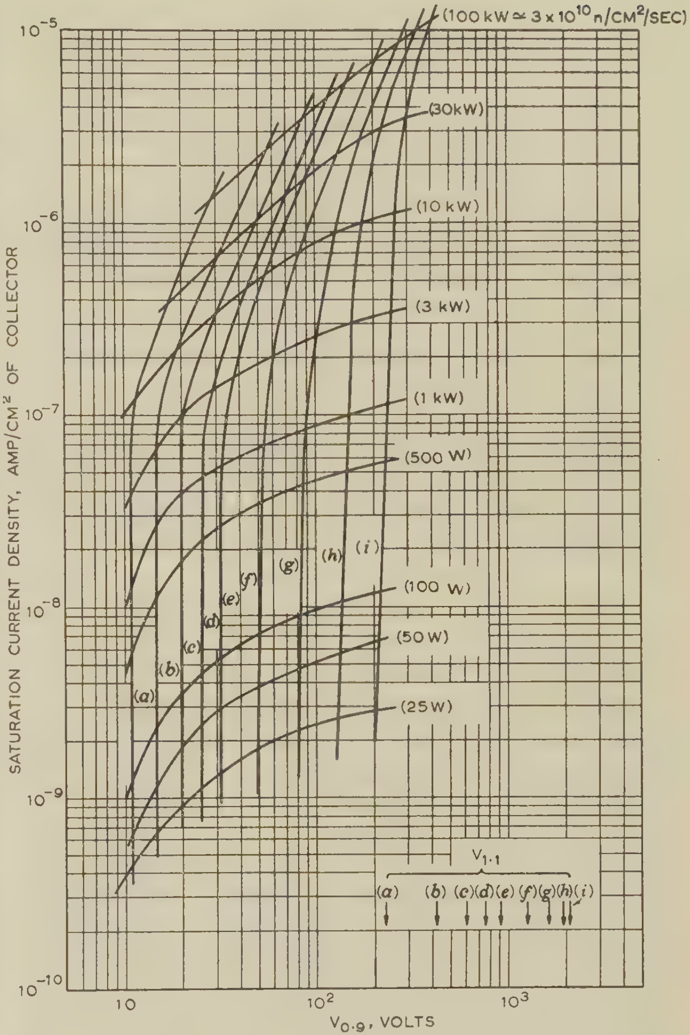


Fig. 2.—Variation of $V_{0\cdot9}$ with saturation current density and p for xenon and an electrode spacing of 1 mm.

(a) 140 mm Hg. (d) 605 mm Hg. (g) 1 520 mm Hg.
(b) 300 mm Hg. (e) 760 mm Hg. (h) 1 900 mm Hg.
(c) 450 mm Hg. (f) 1 140 mm Hg. (i) 2 280 mm Hg.

values, except in neon, which is very sensitive to traces of argon contamination.

Some general conclusions may be drawn from the experimental results. For any saturation current $V_{1\cdot1}/V_{0\cdot9}$ is greatest in nitrogen and lowest in neon, so that relatively longer plateaux can be obtained in nitrogen and xenon than in neon. For any saturation current $V_{0\cdot9}$ is lowest in neon and greatest in xenon, so that lower operating potentials are possible in neon and nitrogen than in xenon. It is clear that the operating characteristics obtained with nitrogen as a filling gas are superior to those obtained with either neon or xenon. However, it is thought that nitrogen may react chemically with uranium oxide and stainless steel after prolonged exposure at a temperature of 500° C, which might result in damage to the fissile coating. Xenon was therefore chosen in preference to neon because of the better and more reliable saturation characteristic, even though it requires a higher operating potential.

Fig. 3 shows the variation of $V_{1\cdot1}/V_{0\cdot9}$ as a function of gas-filling pressure for xenon in a chamber with an inter-electrode gap of 1 mm, at different saturation current densities. It may

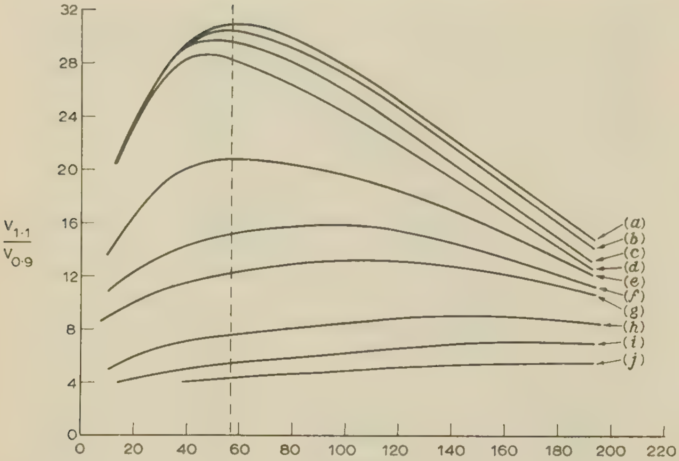


Fig. 3.—Experimental values of $V_{1\cdot1}/V_{0\cdot9}$ as a function of xenon pressure for various saturation current densities, with a 1 mm electrode gap.

Collector current densities, A/cm²

(*a*) 8×10^{-10}. (*b*) 2×10^{-9}. (*c*) 3×10^{-8}. (*d*) 1×10^{-7}. (*e*) 3×10^{-7}. (*f*) 6×10^{-7}. (*g*) 1×10^{-6}. (*h*) 3×10^{-6}. (*i*) 6×10^{-6}. (*j*) 1×10^{-5}.

be seen that at low current densities the ratio increases considerably as p is decreased down to about 60 cm Hg, below which pressure it decreases again. At higher current densities the variation is much smaller, but $V_{1\cdot1}/V_{0\cdot9}$ tends to decrease as p is decreased.

It was also found that $V_{1\cdot1}/V_{0\cdot9}$ increases as the inter-electrode gap, d, decreases, and thus, except at the highest current densities, a reduction of both p and d leads to relatively longer plateaux. This also has the effect of reducing the operating potential and the γ-sensitivity, which is proportional to the mass of gas contained in the sensitive volume; the accompanying reduction in neutron sensitivity can be compensated for by increasing the quantity of uranium 235 on the electrodes. A gas pressure of 57 cm Hg was chosen as a compromise to maintain an adequate value for $V_{1\cdot1}/V_{0\cdot9}$ without unduly increasing the operating potential.

A lower limit is set to the electrode spacing by the need to place reasonable tolerances on the dimensions of both the electrodes and the ceramic insulators. A standard engineering tolerance of $\pm 0\cdot001$ in on the diameter of each electrode will result in a possible variation of $\pm 10\%$ in the electrode spacing between different chambers if the nominal spacing is 0·02 in (0·5 mm), producing a similar variation in the current sensitivities. Difficulties are also encountered with very small spacings due to the intermittent electrical breakdown caused by small irregularities on the coated surfaces. Accordingly, an electrode spacing of 1 mm and a gas filling of xenon at 57 cm Hg were adopted. Under these conditions a current density of 4×10^{-5} A/cm² can then be collected with an acceptable characteristic showing a well-defined saturation region. $V_{0\cdot9}$ is 250 V, $V_{1\cdot1}$ is 750 V, and the normal operating voltage is 450 V. A maximum current of 1 mA may then be drawn from a chamber in which the collector electrode has an area of 25 cm².

(3) ELECTRICAL CONNECTIONS

The connecting cable for an in-pile ionization chamber must be flexible and robust, must have an adequate electrical insulation resistance at 500° C and must not introduce into the reactor any materials which would be incompatible with either th graphite moderator or the Magnox cladding of the fuel elements This led to the choice of a cable with a sheath of nickel-chrom steel 0·125 in diameter and 0·012 in thick containing a singl nickel-chromium alloy centre conductor and an insulant o closely packed magnesium oxide. The sheath may be argon-ar welded through stainless-steel plates to make vacuum-tight seal without affecting the electrical insulation. It may be woun round a drum of 18 in diameter and does not easily fracture If the ends of the cable are baked thoroughly before sealing of a conductance of less than 10^{-14} mho/ft can be maintaine between the inner and outer conductors at normal ambien temperature, which will increase to 2×10^{-13} mho/ft at 250° C and to 2×10^{-10} mho/ft at 500° C, when a potential differenc of at least 750 V can be applied between the inner and oute conductors before breakdown occurs. Under irradiation in th core of a graphite-moderated reactor at a neutron flux intensit of 10^{13} n/cm²/sec and a γ-radiation dose-rate of about 10^7 r/h the apparent conductance increases to about $1\cdot4 \times 10^{-10}$ mho/f at 250° C. For these radiation levels the radiation-induced con ductivity is the limiting factor governing spurious cable current for temperatures up to 250° C. At higher temperatures th intrinsic conductivity increases and becomes the dominant facto

The leakage currents measured are approximately propor tional to applied voltage and to radiation intensity, but the may exhibit long relaxation times. The time required to achiev a steady-state condition after a change in applied voltage o radiation intensity may be of the order of one hour; the con ductances quoted above correspond to steady-state conditio There is no evidence of any permanent deterioration in electrica characteristics after an irradiation period of several months in flux intensity of 10^{13} n/cm²/sec.

A maximum saturation current of 1 mA may be drawn fro a chamber in which the collector electrode has an area of 25 cm and by adjusting the weight of uranium in the fissile coating this current may be drawn at the maximum design flux o 10^{13} n/cm²/sec. The chamber must operate satisfactorily a lower flux intensities to accommodate variations in the flu pattern through the reactor core at full power and at reduce powers. If this range is taken as 30 : 1, comprising a 3 : spatial variation and a 10 : 1 power variation, the minimu output current from the chamber will be 30 μA, and to preserv adequate linearity the leakage current due to chamber insulato

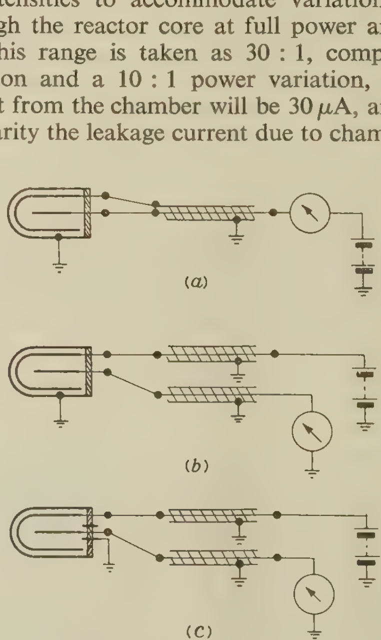


Fig. 4.—Methods of making external connections to an ionizatio chamber.

(*a*) Single connecting cable with non-guard-ringed chamber.
(*b*) Guard-ringed collector cable with non-guard-ringed chamber.
(*c*) Fully guard-ringed chamber and collector cable.

and cables should be less than 3% of this value, i.e. 1 μA. At 500° C a leakage current of about 10 μA will be obtained from 100 ft of cable with 500 V applied between its inner and outer conductors. The appearance of this current as part of the chamber output can be prevented by using two connecting cables to the chamber as shown in Fig. 4(*b*), since the potential difference applied across the collector cable is only that developed across the input of the measuring instrument. A similar reduction in the effect of electrode leakage current may be achieved by the use of guard-ring electrodes [Fig. 4(*c*)]. This arrangement was adopted because it allows a greater margin of safety in the design and also permits a uniform electric field intensity to be maintained across the collecting volume of the chamber.

(4) CHAMBER CONSTRUCTION

The chamber, shown in Fig. 1, embodies a cylindrical electrode structure in which the outer cylinder is the h.t. electrode and the inner is the current or collector electrode; guard-ring electrodes are included at each end of the collector. Both electrodes are coated with fissile material to prevent any net transfer of uranium from one electrode to the other by evaporation during the fission process.[11] The coated area on each electrode is 25 cm², and the fissile coating is applied as a solution of uranium nitrate which is then oxidized to uranium oxide by baking in air.[12]

The metal parts of the chamber are manufactured from 18/8/1 stainless steel and all joints are argon-arc welded and both vacuum and pressure tested. Electrical connections through the walls of the chamber to the electrodes are made through alumina insulators which are joined to a suitable metal flange by a vacuum-tight braze. The internal insulators supporting the electrodes are accurately ground from alumina of 99·5% purity, and the insulation resistance at 500° C between the h.t. electrode and the guard rings, and between these and the collector electrode, is greater than $10^8\,\Omega$. The insulators between the guard rings and the collector are shielded from the fissile coating to minimize any deposition of uranium on them by the evaporation process mentioned above.

The outer parts of the metal-alumina seals are protected by enclosure in a separate vacuum-tight compartment, which is filled with xenon to the same pressure as in the ionization chamber, in order to protect the braze materials from oxidation at high temperatures and to minimize the effect of any gas leaks. An electrostatic shield is placed round the collector seal to prevent the collection of any ions formed in the seal compartment. For prototype tests in a Calder Hall type of reactor[13] an outer sealed containment envelope (not shown in the illustration) provided a further protection against possible contamination of the reactor core with fission products from the uranium 235 coating in the event of failure of the primary seals or joints.

Care is necessary during assembly of the chambers if the characteristics are to remain stable for extended periods during operation. Metal parts are outgassed at 800° C for several hours and the internal insulators are baked in air at 1000° C. After an ionization-chamber compartment has been assembled it is baked at 500° C with an air filling to remove any trace of cleaning materials or dust particles and then evacuated to less than 10^{-5} mm Hg until the apparent leakage rate at 500° C is below 5 litre-microns in one hour. The chamber is then cooled under vacuum and the filling gas is admitted. A similar procedure is followed before filling the seal compartment, except that, owing to the release of gas from the mineral-insulated cables, the apparent leakage rate is considerably higher. When the chamber is completed a final test is made of the insulation resistance at 500° C.

(5) RESPONSE IN A GRAPHITE-MODERATED REACTOR

When a chamber is irradiated in a reactor the most important fraction of the output current is that which is immediately associated with the neutron flux. This 'prompt' component is due to ionization from fission fragments in the chamber, from the effect of prompt γ-radiation produced during fission in the reactor fuel and by neutron absorption in the materials of the reactor and ionization chamber. The remaining component of the chamber current, although ultimately proportional to the neutron flux, is not immediately associated with it, and because it tends to persist after the removal of the neutron flux it is known as the 'residual' current. It is due to the β- and γ-radiations from fission products and activated materials in the chamber and the reactor, the major contribution being β-activities in the chamber and γ-activities from the reactor.

In principle, the various contributions to the output current can be calculated. That due to fission fragments from the coating is obtained from their known ranges and rate of energy loss[14] and from the electrode geometry. To calculate the contributions from β- and γ-radiations use is made of the Bragg–Gray principle,[15] namely that, in a small gas volume surrounded by a solid medium,

$$W_s = \frac{S_s}{S_g} W_g$$

where W_s and W_g are the energy absorptions per unit mass of solid and gas respectively, and the mass stopping powers for β-radiation are S_s and S_g. W_g can be replaced by $N_g w$, where N_g is the number of ion pairs formed per unit mass of gas and w is the mean energy required to produce one ion pair in the gas.

S_s/S_g is very near to unity, and hence $W_s \simeq N_g w$. The value of W_s for γ-radiation, whether prompt or residual, is found by using suitable build-up factors and integrating over the entire reactor system and the detector to obtain the dose-rate at the chamber electrodes. The contribution from β-activation of the chamber materials is obtained by calculating W_s for the β-decay of each radioactive isotope formed by neutron absorption; with stainless-steel electrodes the main contribution is produced by the 1–2% manganese which is present. To estimate the effect of β-radiations from fission products released in the chamber a mean free path for the β-particles in the gas is calculated and an average value of the total specific ionization assumed.

The calculations are necessarily inexact, because of the simplifying assumptions which it is convenient to make, and also because the data which form the basis of the calculations are insufficiently precise. This is particularly so for the prompt γ-radiation emitted after neutron absorption.

A series of experiments were therefore conducted in the Bepo reactor with models of the basic ionization chamber, some of which had stainless-steel electrodes coated with $^{235}U_3O_8$ and others with uncoated electrodes of stainless steel or titanium. The chambers were irradiated in a neutron flux intensity of $1{\cdot}3 \times 10^{12}$ n/cm²/sec for periods of 2–3 weeks to allow a build-up of the longer-lived activities, although the effect of any small amount of cobalt impurity could not be measured after such short irradiations. A stainless-steel shell which could be placed round one of the chambers with uncoated steel electrodes was used to distinguish the effects of prompt γ-radiation due to the reactor materials from those of the chamber.

The output currents from the chambers were measured at full power and for some time after a reactor shut-down, both in and out of the reactor. By removing the chambers from the reactor a short time after shut-down the residual currents due to the chamber activities could be measured, and by comparing the different results from chambers having stainless-steel and titanium electrodes the currents due to β- and γ-activities could

be separated. The decay of the residual current in both coated and uncoated chambers was followed for periods of 1 000 h after they had been removed from the reactor. In the chambers with uncoated electrodes the main component of residual current had an initial half-life of 2·57 h, corresponding to the decay of manganese 56, and later this became 27·8 days, corresponding to that of chromium 57. By subtracting the residual current of an uncoated chamber from that of one containing fissile material the contribution of the fission product activities could be deduced, and by extrapolating these residual currents to the time of shut-down and comparing them with the current measured when the chambers were left in the reactor, the contribution due to the reactor activities was obtained.

Table 1

CURRENTS FROM IONIZATION CHAMBERS IN GRAPHITE-MODERATED REACTOR

Source of current	Relative currents: Coated stainless-steel electrodes	Plain stainless-steel electrodes	Plain titanium electrodes
	%	%	%
Prompt radiations			
Fission in uranium 235 coating	97	0	0
n, γ reactions in reactor core . .	2·4 (both rows)	66·8	95·4 (both rows)
n, γ reactions in chamber materials		24·5	
Total prompt effects . .	99·4	91·3	95·4
Residual radiations			
Fission-product activity in uranium 235 coating	0·2	0	0
Fission-product activity from reactor core	0·28	3·8	3·4
Induced β-activity in chamber	0·12 (both rows)	3·8	0
Induced γ-activity in chamber		1·1	1·2
Total residual effects	0·6	8·7	4·6
Ionization-chamber data			
Total electrode area, cm² . .	50	50	50
Weight of uranium 235 on electrode, mg	0·48	0	0
Electrode spacing, mm . .	1	1	1
Pressure of xenon, cm Hg . .	57	220	220
Thermal-neutron flux intensity,	10^{13}	10^{13}	10^{13}
Total measured current, mA . .	0·7	0·14	0·13

Table 1 compares the experimental output currents of chambers with various electrodes, and Figs. 5 and 6 show typical decay curves of the residual currents in and out of the reactor after shut-down. All the results have been adjusted to correspond to an irradiation period of 1 year in a neutron flux intensity of 10^{13} n/cm²/sec.

The prompt current in a chamber with uncoated steel electrodes is about 90% of the total output, the remaining 10% being due to residual effects. The exact proportion of the current due to residual effects depends on the amount of manganese present in the steel and on the total quantity of steel used in the construction. About 40% of the residual current is due to induced β-activities in the electrodes and the remainder to γ-activities in the reactor and chamber. The effect of β-activity in the electrodes can be reduced by making the electrodes of a material which has a low neutron-activation cross-section, such as titanium, when the prompt component of current increases to 95% of the total output; 75% of the residual current is then due to fission product activity in the reactor core.

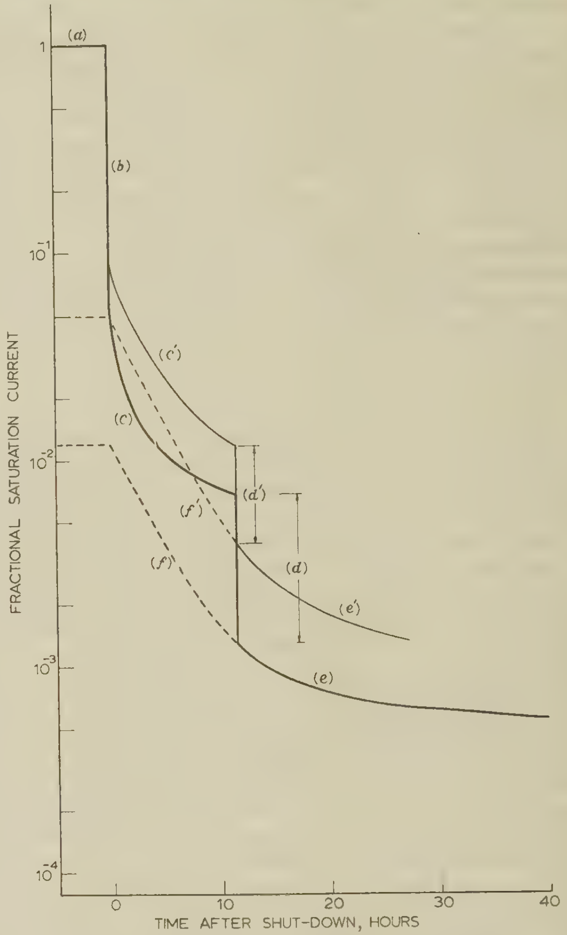


Fig. 5.—Saturation currents of chambers with plain stainless-steel and titanium electrodes after reactor shut-down.

Electrode spacing: 1 mm.
Gas filling: 220 cm Hg of xenon.
Neutron flux intensity: 10^{13} n/cm²/sec.

(*a*) Equilibrium saturation currents: 0·14 mA for stainless steel and 0·13 mA for titanium.
(*b*) Reactor shut down.
(*c*), (*c′*) Total residual activity in reactor for titanium and stainless steel respectively.
(*d*), (*d′*) Chambers removed from reactor.
(*e*), (*e′*) Induced activities in chambers for titanium and stainless steel respectively.
(*f*), (*f′*) Induced activities extrapolated back to reactor shut-down for titanium and stainless steel respectively; values at $t = 0$ are 1·55 and 7·0 μA respectively.

Table 2

RESIDUAL CURRENT FROM A CHAMBER IN THE REACTOR AT VARIOUS TIMES AFTER SHUT-DOWN* AS A PERCENTAGE OF FULL POWER CURRENT

Time after shut-down	Total residual current: Plain stainless-steel electrodes	Plain titanium electrodes	Uranium 235-coated electrodes
hours	%	%	%
0	9	4·6	0·6
2	4·95	1·85	0·25
5	2·6	1·06	0·125
10	1·3	0·71	0·07
15	0·88	0·6	0·05

* The reactor is assumed to shut-down instantaneously.

The total current delivered by the chamber with uranium-coated electrodes was 0·7 mA in a neutron flux intensity of 10^{13} n/cm^2/sec, i.e. somewhat less than the maximum current of 1 mA quoted in Section 2. In this chamber the current due to residual effects is reduced to only 0·6% of the total output, 0·4% being due to reactor fission products and induced activities in the chamber and the remaining 0·2% to fission product activities released from the coating. The latter figure represents an irreducible residual component which limits the operating range of this type of chamber.

The rate at which the residual current in the coated and uncoated chamber decays when the chambers are left in the reactor after an instantaneous shut-down is shown in Table 2, from which it can be seen that the chamber with uranium-coated electrodes shows the fastest decay, and the chamber with plain steel electrodes shows the slowest.

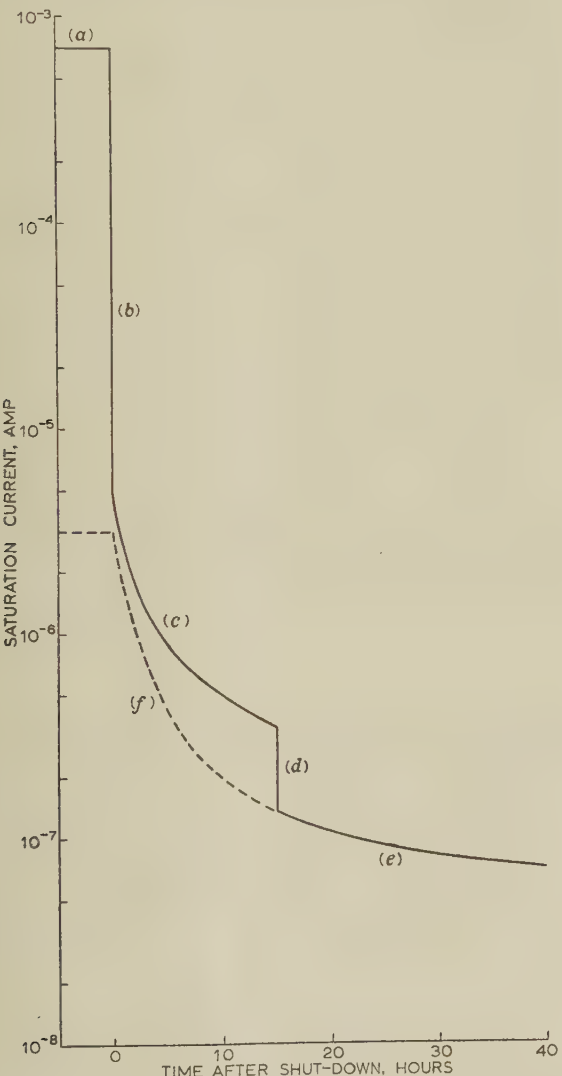


Fig. 6.—Saturation current of uranium 235 coated chamber after reactor shut-down.

Electrode spacing: 1 mm.
Gas filling: 57 cm Hg of xenon.
Neutron flux intensity: 10^{13} n/cm^2/sec.
Total fissile coating: 0·48 mg of uranium 235.

(a) Equilibrium saturation current = 0·7 mA.
(b) Reactor shut-down.
(c) Total residual activity in reactor.
(d) Chamber removed from reactor.
(e) Induced activities and fission product activities in chamber.
(f) Curve (e) extrapolated back to reactor shut-down.

(6) DEPLETION EFFECTS IN CHAMBERS WITH COATED ELECTRODES

All long-term irradiation tests so far carried out with chambers whose electrodes are coated with uranium 235 have shown that depletion of uranium due to the fission process is the major factor contributing to a change of sensitivity with time. The calculated reduction in sensitivity from this effect is 20% in 1 year in a neutron flux of 10^{13} n/cm^2/sec. This may be undesirable and can be minimized by using a coating which contains a mixture of uranium 235 and uranium 238. As a result of neutron capture in the isotope 238 the nuclide plutonium 239 is formed by the process

$$^{238}\mathrm{U}\,(n, \gamma)\,^{239}\mathrm{U} \xrightarrow{\beta 23\cdot 5\,\mathrm{min}} {}^{239}\mathrm{Np} \xrightarrow{\beta 2\cdot 33\,\mathrm{d.}} {}^{239}\mathrm{Pu}$$

Plutonium 239 is fissionable by thermal neutrons and its formation can be made to compensate for the depletion of uranium 235.

If the ratio of the number of uranium 238 atoms to those of uranium 235 is initially R, it may be shown that the average fission cross-section, $\bar{\sigma}_f$, per atom of the mixture after a time t in a neutron flux intensity ϕ is

$$\bar{\sigma}_f = \sigma_{f5} \exp(-\sigma_{a5}\phi t) + \frac{R\sigma_{a8}\sigma_{f9}}{\sigma_{a9}} \times \left(1 - \frac{\sigma_{a8}}{\sigma_{a9}}\right)^{-1} [\exp(-\sigma_{a8}\phi t) - \exp(-\sigma_{a9}\phi t)]$$

It is assumed that above thermal energies the neutron energy spectrum is similar to that in a Bepo-type reactor (graphite moderated). The neutron flux and the cross-sections are defined according to the Westcott convention.[16]

If the average energy expended per fission in the chamber gas is the same for fission fragments from plutonium 239 as from uranium 235, the current sensitivity will follow the change in $\bar{\sigma}_f$, which in general increases for short irradiation times and then decreases. The fission and absorption cross-sections as defined above are dependent on the incident neutron energy, and hence the best value of R for any application will depend on the temperature of the moderator in the vicinity of the ionization chamber. Cross-sections which are appropriate for the moderator of a reactor of the Calder Hall type, and which were used to determine R, are given in Table 3.

Table 3

EFFECTIVE CROSS-SECTIONS OF URANIUM 235, URANIUM 238 AND PLUTONIUM 239 IN A LARGE GRAPHITE-MODERATED REACTOR

Temperature	σ_{a5}	σ_{a8}	σ_{a9}	σ_{f5}	σ_{f9}
deg K	barns	barns	barns	barns	barns
300	665	20·9	1 223	559	841
400	653	21·8	1 360	550	909
500	644	22·4	1 559	542	1 018
600	638	22·9	1 820	536	1 162
700	635	23·3	2 117	533	1 330
800	632	23·7	2 417	530	1 503

Fig. 7 shows the calculated percentage change in $\bar{\sigma}_f$ of a mixture containing a 14 : 1 atomic ratio of uranium 238/uranium 235 in a thermal-neutron flux intensity of 10^{13} n/cm^2/sec at various moderator temperatures. This ratio shows a maximum variation during one year of ±3% for any temperature in the range 300–500° C. If the sensitivity of the chamber is to be the same as that described in Section 5, the electrode coating will contain 0·48 mg of uranium 235 and 6·25 mg of uranium 238.

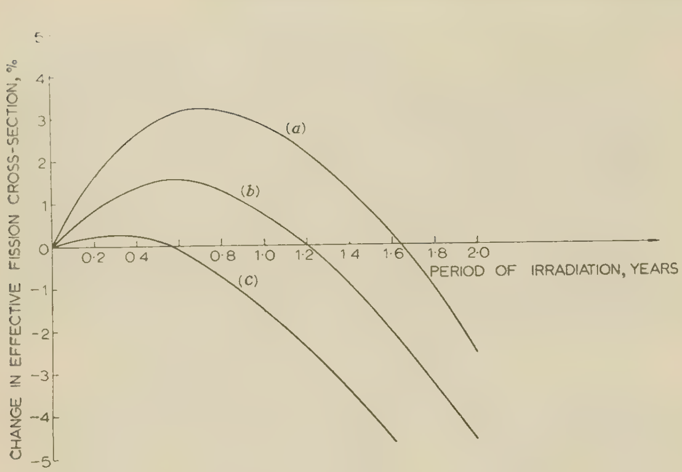


Fig. 7.—Calculated depletion effects in 14 : 1 uranium 238/235 mixture.

Thermal neutron flux intensity: 10^{13} n/cm²/sec.

(a) 800° K.
(b) 700° K.
(c) 600° K.

(7) OPERATIONAL AND ACCELERATED LIFE TESTS

Several chambers have been irradiated in a Westcott flux of $3 \cdot 8 \times 10^{13}$ n/cm²/sec in the Pluto reactor, to study the effects of radiation on the saturation characteristics and current sensitivity and also to examine the effects of depletion of the fissile coating.

One chamber containing 0·92 mg of uranium 235 (coating thickness, 0·018 mg/cm²) was operated at 320° C for 2200 h, equivalent to an irradiation for one year in a flux of 10^{13} n/cm²/sec. The current sensitivity of this chamber was initially $2 \cdot 16 \times 10^{-16}$ A/n/cm²/sec. No marked change in the saturation characteristics was observed, but the sensitivity was reduced by 20%, compared with a calculated value of 17%. This discrepancy may have been due to errors in measurement of the integrated neutron flux over the irradiation period.

Since radiation damage to the fissile coating might be more significant for thicker coatings, a chamber containing 6·47 mg of uranium (coating thickness, 0·13 mg/cm²) was irradiated at 350° C. This coating weight is about the same as that proposed in Section 6 for a chamber in which depletion effects are minimized, but the ratio of uranium 238/235 was reduced to 6 : 1 so that the total number of fission events occurring in 1 000 h would be the same as would occur in 1 year at 10^{13} n/cm²/sec. No change in the characteristics was observed, apart from some decrease in the sensitivity due to depletion effects. Other chambers have been irradiated for periods of up to 2 000 h at temperatures up to 450° C in a neutron flux of $3 \cdot 8 \times 10^{13}$ n/cm²/sec without appreciable deterioration. Since the neutron-energy spectrum in the Pluto reactor differs from that in a graphite-moderated reactor, it has not been possible to test quantitatively the depletion effects in uranium 238/235 mixtures, but there is evidence that the current sensitivity of chambers containing mixed coatings falls less rapidly than for pure uranium 235 coatings.

Three chambers have been irradiated together in a Calder Hall type of reactor for 10 months at an average flux intensity of $3 \cdot 5 \times 10^{12}$ n/cm²/sec and a temperature of 260° C to gain experience in inserting them under operational conditions into a large graphite-moderated reactor; no appreciable change in characteristics has so far been observed. A further three chambers were recently loaded into a similar reactor at a neutron flux of 10^{13} n/cm²/sec and a temperature of 260° C, and will remain in the reactor for at least 9 months. The electrodes of one of these chambers was left uncoated, and it will provide further information concerning the fraction of total residual current due to the build-up of activities with very long half-lives, such as cobalt 60. The other two chambers contain mixed coatings with uranium 238/235 ratios of 16 : 1 and 12 : 1 and will provide data on the extent to which the build-up of plutonium 239 compensates for the depletion of uranium 235. This programme will be complete in about 1 year.

(8) CONCLUSIONS

It has been shown that mean-current fission ionization chambers can be made to operate successfully for a total radiation dose of 3×10^{20} n/cm² at temperatures up to 450°C.

For adequate measurement and control of spatial instabilities in a reactor neutron flux distribution it is necessary that small prompt changes in this distribution as measured by an in-pile ionization chamber should not be obscured by changes in the output current due to the build-up or decay of the residual component of current. Spatial instabilities in the flux distribution would be detected by relative changes in the signals from several chambers, so that if the residual currents from all chambers varied in the same manner the effect would not be serious. Nevertheless, if the residual current produces a change of more than 5% in the output signal when it is required to detect relative changes of 1% between chambers, the performance cannot be regarded as satisfactory.

To decide whether a chamber is suitable for flux-instability control it is necessary to inquire over what range the reactor power level may be altered before the subsequent change in the output signal reaches 5% due to the build-up or decay of the residual current component. For a chamber with coated electrodes designed to deliver a current of 7×10^{-4} A in a neutron flux intensity of 10^{13} n/cm²/sec, i.e. a current density of $2 \cdot 8 \times 10^{-18}$ A/cm² of collector electrode per n/cm²/sec, the residual current in equilibrium is 0·6% of the total, so that any small change in the flux distribution following an overall increase in the neutron flux intensity can be detected satisfactorily. The flux intensity cannot rapidly be reduced to less than 0·1 of a previous operating level since otherwise the subsequent change in the residual component would be greater than 5% of the total output.

The performance of a chamber with uncoated steel electrodes is less satisfactory because the residual current component in equilibrium is about 10% of the output. This limits the allowable rapid increase in neutron flux to about twice a previous value and a rapid decrease to 0·66 of a former value if the subsequent change in the residual component is not to exceed 5% of the new output current. Some improvement may be achieved by a chamber with titanium electrodes, in which the residual component is only 4·6% of the total. Any increase in neutron flux intensity can be allowed, and the allowable decrease is limited to 0·45 of the former value.

The maximum current density that may be collected from the present design of fission ionization chamber is about 4×10^{-5} A/cm² if an acceptable saturation characteristic is to be maintained. This output may be achieved at different flux intensities by adjustment of the amount of uranium 235 in the coating, but a lower limit of 10^{11} n/cm²/sec is set even with a pure uranium 235 coating by the need to restrict the total thickness to 1 mg/cm²; this is the maximum thickness that may be applied with good adherence. In this flux intensity the equilibrium residual component due to chamber and reactor activities would account for only 0·004% of the total output, though fission product activity in the chamber would still be 0·2%. This latter effect places a fundamental limit on the operating

range of the chamber, and it is doubtful whether it could be greatly improved.

If the coating thickness is adjusted to deliver the maximum current in a flux intensity of 10^{14} n/cm^2/sec, the equilibrium residual current due to the chamber and reactor activities will rise to 4% of the total output, while the component due to fission-product activity in the chamber will remain at 0·2%. The chamber will still satisfactorily cope with any increase in flux intensity, but a decrease would now be limited to 0·45 of its previous value.

Thus the fission ionization chamber has a much greater operating range than one with plain titanium electrodes at flux intensities below 10^{13} n/cm^2/sec, but above this level the differences become progressively less marked until, at 10^{14} n/cm^2/sec, the residual components of the two types are about the same fraction of the total current. At this level the chamber with titanium electrodes is more attractive, since it will not be subject to changes in current sensitivity associated with depletion of a fissile coating. This type of chamber may also have an advantage at flux intensities as low as 10^{13} n/cm^2/sec when very long irradiation periods are required.

The equilibrium current sensitivity of such a chamber filled with xenon at 220 cm Hg and having a 1 mm inter-electrode gap will be approximately $5 \cdot 6 \times 10^{-19}$ A/cm^2 of collector electrode per neutron per square centimetre per second, and at a flux intensity of 10^{13} n/cm^2/sec the operating potential will be 300 V. At higher flux intensities it will be necessary to reduce the current sensitivity by reducing the gas pressure in order to maintain an acceptable saturation characteristic.

In flux-scanning applications a chamber need be exposed to a high flux intensity only for infrequent and short periods of time. It is then possible that the chamber activities would not build up to significant levels, and a linear response could be achieved over a wider range of neutron flux intensity. It is unlikely that a plain-electrode chamber would be useful in this application, since it is not certain that the prompt component of the current would be exactly proportional to the neutron flux intensity, owing either to the possibility of γ-radiation 'streaming' up the flux-scanning hole in the moderator or to the possibility of changes in the energy spectrum of the γ-radiation across the reactor core.

(9) ACKNOWLEDGMENTS

Our thanks are due to many colleagues for their advice and co-operation in the course of this investigation, but in particular we are grateful to Mr. G. Knill and Mr. R. P. Henderson for their able assistance at all stages of the experimental work. We acknowledge the help given by members of Engineering Division who constructed the various ionization chambers, and to others in the Research Reactor Division at A.E.R.E. and in the U.K.A.E.A. Industrial Group at Chapel Cross who assisted in the operational tests.

(10) REFERENCES

(1) ABSON, W., and WADE, F.: 'Nuclear-Reactor-Control Ionization Chambers', *Proceedings I.E.E.*, Paper No. 2029 M, September, 1956 (**103** B, p. 590).

(2) GRAY, A. L., and WEBSTER, R.: 'Pulse Transformers for Counting', *Nuclear Power*, 1959, **4**, p. 106.

(3) JACQUES, T. A., BALLINGER, H. A., and WADE, F.: 'Neutron Detectors for Reactor Instrumentation', *Proceedings I.E.E.*, Paper No. 1433, May, 1953 (**100**, Part 1, p. 110).

(4) GUS'KOV, YU, K., and ZVONAREV, A. V.: 'Thermocouple System for Measurement of Large Neutron Fluxes', *Instruments and Experimental Techniques*, 1959, No. 5, p. 821.

(5) KOCK, L., LABEYRIE, J., and TARESSENKO, S.: 'New Method of Measuring the Neutron Fluxes in Atomic Reactors', Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy, 1958, Volume 11, Reactor Safety and Control, p. 512.

(6) STUBBS, F. S.: Private communication.

(7) ABSON, W., and AWCOCK, M.: 'Two Ionization Chambers for Activity Measurements on Wires used for Neutron-Flux Scanning', Atomic Energy Research Establishment Memorandum, AERE EL/M. 100, September, 1958.

(8) BOAG, J. W., and WILSON, T.: 'The Saturation Curve at High Ionization Intensity', *British Journal of Applied Physics*, 1952, **3**, p. 222.

(9) DENNIS, J. A., and LOOSEMORE, W. R.: 'The Saturation Characteristics of Mean-Current Fission Ionization Chambers', Atomic Energy Research Establishment Memorandum, AERE M.566, October, 1960.

(10) MEEK, J. M., and CRAGGS, J. D.: 'Electrical Breakdown of Gases' (University Press, Oxford, 1953).

(11) ERSHLER, B. V., and LAPTERA, F. S.: 'The Evaporation of Metals by Fission Fragments', *Journal of Nuclear Energy*, 1957, **4**, Part II, p. 471.

(12) LOOSEMORE, W. R.: 'Acceptance Test Schedule for Pulse-type Fission Ionisation Chambers', Atomic Energy Research Establishment Memorandum, AERE M.728, October, 1960.

(13) LOOSEMORE, W. R., and DENNIS, J. A.: 'An Ionisation Chamber for the Continuous Measurement of Thermal Neutron Flux Intensity at High Temperature', Atomic Energy Research Establishment Report AERE-R. 3474.

(14) LASSEN, M. O.: 'On the Energy Loss by Fission Fragments Along their Range', *Det kgl. Danske Videnskabernes Selskab Matematisk Fysiske Meddelelser*, 1949, **25**, No. 11.

(15) GRAY, L. H.: 'Radiation Dosimetry, Parts I and II'. *British Journal of Radiology*, 1937, **10**, pp. 600 and 721.

(16) CAMPBELL, C. G., and FREEMANTLE, R. G.: 'Effective Cross Section Data for Thermal Reactor Calculations', Atomic Energy Research Establishment Report, AERE RP/R. 2031, August, 1956.

681.142 : 621.317.79 : 621.391

The Institution of Electrical Engineers
Paper No. 3270 M
July 1960

A UNIVERSAL NON-LINEAR FILTER, PREDICTOR AND SIMULATOR WHICH OPTIMIZES ITSELF BY A LEARNING PROCESS

By Professor D. GABOR, F.R.S., Member, W. P. L. WILBY, Ph.D., and R. WOODCOCK, Ph.D.

(*The paper was first received* 17*th October*, 1959, *and in revised form* 22*nd March*, 1960. *It was published in July*, 1960, *and was read before the* Measurement and Control Section 6*th December*, 1960, *the* North-Western Measurement Group 17*th January*, *and the* Rugby Sub-Centre 1*st February*, 1961.)

SUMMARY

A machine is described consisting of a universal non-linear filter, which is a highly adaptable analogue computer, together with a training device. The analogue machine has 18 input quantities from which it can compute in about 2·5 millisec 94 terms of a polynomial, each term containing products and powers of the input quantities, with adjustable coefficients, and can form their sum. The input quantities may be, for instance, 18 past samples of the values of a stochastic variable which is fed into the machine, and the result of the computation is an output function which contains 94 free variables. The training device optimizes the output by successive adjustment of the variable coefficients, until it has approached a target function as closely as can be achieved with a polynomial of 94 terms, by the criterion of the least mean-square error. This is done by repeatedly feeding into the machine a record of the stochastic process, long enough to be representative, and adjusting the variable coefficients, one at a time after each run, by a strategy which ensures that the error will monotonically decrease from run to run.

In order to make the machine an optimum filter it is trained on a record of a noisy process, together with a target record which contains the signal only. It is taught as a predictor by taking as the target function a value of the stochastic process advanced by a certain time interval beyond the last value which goes into the input. It is trained as a simulator, for instance of an unknown mechanism, by feeding it with the input of the mechanism to be simulated at one end and presenting it at the other with its output as target function. The machine will then make itself into a model of the device to be simulated and the non-linear transfer function of the device can be read off from the final setting of the coefficients, as nearly as it can be represented by a 94-term polynomial. The machine is not confined to single-input systems.

The machine incorporates 80 analogue multipliers of a novel 'piezomagnetic' type which, in its present form, can perform over 1 000 multiplications per second with an error of 0·5% or less.

A few examples of the first test applications of the machine are described.

(1) INTRODUCTION

The problem of designing an optimum linear filter and predictor was first conceived by Kolmogoroff in 1942[1] and shortly afterwards was solved by Wiener.[2] Though Wiener's interest almost immediately turned to non-linear processing[3] and though he devoted much brilliant work to it,[4] the problem of an optimum non-linear filter did not prove mathematically tractable. Wiener and his collaborators Lee and Bose,[5] as well as Zadeh,[6] Singleton[7] and White[8] and many others, have done much to clarify the nature of the problem, but it has been clear for some time that, even if a formal solution could be found, it would not be of much practical use, not so much because of the computational work which it would require, but because of the prohibitive amount of labour involved in collecting the material which might form the basis of such a computation.* On the other hand the problem never ceased to be attractive. One can assess what a non-linear filter backed by a large memory can do, by noting that our ear-plus-brain is such a filter, which, as everybody knows, can perform near-miracles in separating music or conversation from a noisy background. The prediction of complicated stochastic processes may well be one of the most important contributions of electronics to the solution of problems of great human importance in economics, sociology and other fields.

One of us proposed[9] in 1954 to take a short cut through the mathematical difficulties by constructing a filter which optimizes its response as do animals and men—by learning. The filter itself must be the realization of a highly flexible mathematical operator, capable of operating on the present and past values of any time function which is fed into it. The flexibility is provided by building into it, say, 100 adjustable parameters. In the teaching process samples of the type of stochastic series which it is to filter or to predict are fed into the machine, together with a target function which is the time-function which the machine is expected to produce. The samples must be long enough to be representative. As an example, the input may be signal plus noise and the target function the signal alone, or the input can be any stochastic series and the target its future value, say one minute later. The instruction to the machine is to adjust its parameters so as to approximate its output as closely as possible to the target function. This is done by running the same sample repeatedly through the machine, adjusting the parameters one by one or several at a time, so as to reduce the error.

There are two conditions for such a machine to be successful. One is that it must have sufficient perceptive power to sense and to store the statistical regularities of the stochastic process on which it is to operate. The second is that there must exist only one optimum solution, so that the machine cannot get stuck at a subsidiary minimum. One can add as a third that the machine must not wear out before it has learned its job.

The range of applications of such a machine goes far beyond what is suggested at first sight by the word 'filter'. This engineering term in its widest sense is coextensive with any quantitative operation on live information, i.e. processing the information at the rate at which it comes in. While the application of a filter as a predictor was at once realized by Kolmogoroff and by Wiener, it was probably Zadeh who first saw that a universal filter can also be used as a recognizer. It must be remembered that noise is 'anything which is not wanted' (Parker); hence, if one wants to recognize a certain type of signal one must only instruct the machine to consider everything else as noise. If the machine can single out certain communication signs, it can also

Prof. Gabor is Professor of Applied Electron Physics and Drs. Wilby and Woodcock were formerly in the Department of Electrical Engineering, Imperial College of Science and Technology, University of London.

Dr. Wilby is now with J. F. Crosfield, Ltd., and Dr. Woodcock is with the Marconi's Wireless Telegraph Co., Ltd.

* An interesting attempt has been made by Lubbock [Lubbock, J. K.: 'The Optimization of a Class of Non-Linear Filters', *Proceedings I.E.E.*, Monograph No. 344 E, November, 1959 (107 C, p. 69)] which is remarkable because it leads to explicit analytical results, at least in a few simple cases. Unfortunately this class of non-linear filters completely lacks the 'perceptive power' of which non-linear systems are capable because its operator is the sum of non-linear operations on single samples of the signal, without any cross-terms such as appear in our operator, eqn. (4).

at once recode them into other signals so long as there is a sufficiently unequivocal correspondence between them. This line of thought leads without any logical break to applications such as speech recognizers and even translators, but it must be pointed out that, as one approaches higher functions of human intelligence, the complication of the machine is likely to increase at a prohibitive rate.

The proper applications of the machine are rather those in which human intelligence is patently inadequate. Human perception is not precisely quantitative, i.e. our senses are rather imperfect measuring instruments, and human memory, though enormous, is not very suitable for storing quantitative information (unless it has been fed in through the sensory organs in digital form). If a human being is expected to manipulate more than about three interdependent parameters to obtain an optimum (for instance, in signal/noise of a receiver) in all probability only fatigue will prevent him from running indefinitely in circles. The machine is here at a distinct advantage, not only because it has infinite patience and a quantitative memory, but because it is free from the deeply ingrained human bias that 'we had it better before'.

An application of the learning filter of great potential interest is as a universal simulator. This has been suggested to us by Professor Tustin.[10] There exist systems of great importance, such as large petrol refineries or the British economic system, which for obvious reasons cannot be submitted to experimental tests, however important it may be to know their behaviour in certain contingencies. On the other hand, data on their inputs and outputs are available over long periods. If one puts these into a learning filter with the instruction 'produce the second set out of the first', the machine will make itself into a model of the system and experiments can then be carried out on the model instead of on the system. Moreover, the machine will give a mathematical formula or formulae for the response functions of the system, which might lead to some insight into its unknown mechanism.

The development of our learning filter started in October, 1956, and the first results were obtained in October, 1959, after just three years' work.

(2) THE MATHEMATICAL PRINCIPLE

Much of the difficulty (though also much of the mathematical elegance) of the Wiener–Kolmogoroff theory of filtering is due to the fact that these authors considered signals in an unlimited frequency range. In practice the interest is always on stochastic processes with a limited frequency band. It is well known[10, 11, 12] that in a frequency band, F, and in a time interval, T, there are $2FT$ degrees of freedom; hence any band-limited signal, in a finite time, can be represented by a finite number of parameters. An elegant and simple representation of signals limited to the band $\pm F$ was introduced by Shannon[13] in 1948 in terms of elementary signals of the type

$$u_n = \frac{\sin 2\pi F(t - t_n)}{2\pi F(t - t_n)} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

By the Whittaker–Shannon sampling theorem, the band-limited part of a time function, $f(t)$, called the cardinal function, can then be expressed by the series

$$[f(t)] = \sum_{t_n = -\infty}^{\infty} f(t_n) u_n \quad . \quad . \quad . \quad . \quad . \quad (2)$$

In this series, the coefficients $f(t_n)$ are the samples of the time function taken at the sampling points $t_n = t_0 + n\tau_0$, where τ_0 is the Nyquist interval, $1/2F$. These samples, $2F$ per unit time, are all the data of the process $f(t)$ which are observable in the frequency band F. (One half of the band $\pm F$ is redundant, but the single-sideband representation is appreciably more complicated.)

Shannon and, in particular, Oswald[14] have shown that all linear processing of a band-limited signal consists of linear algebraic transformations of the coefficients of the series, i.e. of the samples. Singleton,[7] White[8] and Gabor[9] have extended this to non-linear transformations. The calculus (really algebra) of the band-limited functions is of extreme simplicity. It suffers only from the blemish that a physical filter could realize the operation $f(t) \rightarrow [f(t)]$ only with an infinite delay. The u-function has a long 'precursor tail'; i.e. in eqn. (2) at time t, samples taken at future instants, $t_n > t$, have appreciable importance. On the other hand, if one cuts off the series at $t_n \leqslant t$ there will be a truncation error.

It is shown in Section 9.1 how this can be remedied to some extent by introducing the responses of physical, realizable filters as interpolation functions instead of the u_n-functions in eqn. (1). By the expansion theorem, an observer beyond a physical filter of bandwidth F will not notice any difference if the input $f(t)$ is replaced by a certain expansion in terms of the time-response functions, $r(t)$, of the filter, spaced by Nyquist intervals. The expansion theorem in turn suffers from the blemish that a physical filter cannot have a sharply limited frequency band, i.e. its frequency response cannot be exactly zero outside a limited band. Thus, the truncation error in time is exchanged for a truncation error in frequency. In practice this error, though it can never be zero, can be made negligibly small. As the interpretation of the coefficients in the expansion theorem is somewhat complicated, it will not be used here, but reference should be made to Section 9.1. We will use instead, for simplicity, the sampling theorem, but in a form adapted to a live signal:

$$[f(t')] = \sum_{n=0}^{n=N} f(t - n\tau_0) u(t' - t + n\tau_0) \quad t' \leqslant t \quad . \quad (3)$$

Here the signal has been sampled at times t, $t - \tau_0$, . . . $t - N\tau_0$, as illustrated in Fig. 1. The expansion has only a

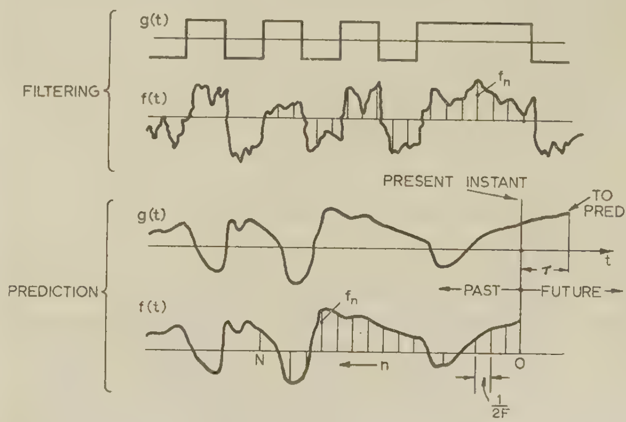


Fig. 1.—Input functions and target functions.

Whether the problem is one of filtering, prediction or simulation, the instruction to the machine during the learning process is always to produce the second from the first, as closely as possible, the success being measured by the criterion of least mean square error.

finite number of terms. A truncation error arises at both ends of the interval (though not of course at the sampling points, which are always exact). It will be assumed that N is large enough to cover the active past, i.e. the interval which is of

appreciable interest for the desired operation of the filter. In the case of prediction, for instance, this is the maximum length of the causal chains contained in the process.

The general problem which the filter has to solve is as follows. There is given a band-limited stationary time series, f(t), i.e. a series with translation-invariant statistical characteristics, from which we have to extract certain information so as to produce a second time series, g(t), to be called the target function. The information is contained in the samples f(t) . . . f($t - N\tau_0$) of the active past, and in certain additional time-independent data. If, for instance, the series to be processed were sea waves, the samples would be taken on the wave profile and the data would be the wind speed, the depth of the sea, etc. We will leave the auxiliary data out of account for the present, as the teaching process could be repeated for every new set of these data. How to deal efficiently with a great variety of auxiliary data, and in particular how to adapt the filter to circumstances when these data are themselves (slowly) changing with time, would lead us into a problem of a higher order than we are prepared to solve at the present stage. We assume, therefore, that we have only the samples of $f_0 \ldots f_N$ to operate on.

It is shown in Section 9.1 that the most general functional of the past of a band-limited time function can be put in the form

$$\mathrm{O}[\mathrm{f}(t)] = \sum_0^N f_n r_n + \sum\sum f_{n1} f_{n2} r_{n1n2} + \sum\sum\sum f_{n1} f_{n2} f_{n3} r_{n1n2n3} + \ldots \quad (4)$$

that is to say in the form of a polynomial of the samples, with certain coefficients. These coefficients are the response functions, defined only for integer values of the arguments, and are therefore better represented by a suffix notation. The first term represents a general linear filter; the r_n factors indicate the weight which a sample f_n, occurring n elementary intervals earlier, is given at the present instant. The second term contains the products of two samples, including their squares; their coefficients, $r_{n_1n_2}$, indicate the weight of a diad at n_1, n_2 at the present instant, and so on. The terms have to cover the active past. It is seen that their number increases steeply with the order. If the active past extends over N intervals, there will be $N + 1$ terms of the first order, $\frac{1}{2}(N + 1)(N + 2)$ terms of the second, $\frac{1}{6}(N + 1)(N + 2)(N + 3)$ of the third, and so on. It is evident that even an operator with very many terms will be able to cover only processes which have a certain perspective, of the type in which the length of the active past of non-linear effects steeply decreases with their order.

The filter is a physical realization of the operator O. It is seen that it has to carry out three types of operation:

(a) Delay the input by 1, 2 . . . N units, i.e. present the samples f_n simultaneously with $f_0 = \mathrm{f}(t)$ to the arithmetical unit.
(b) Multiply these samples with one another.
(c) Add them up, multiplied by adjustable coefficients.

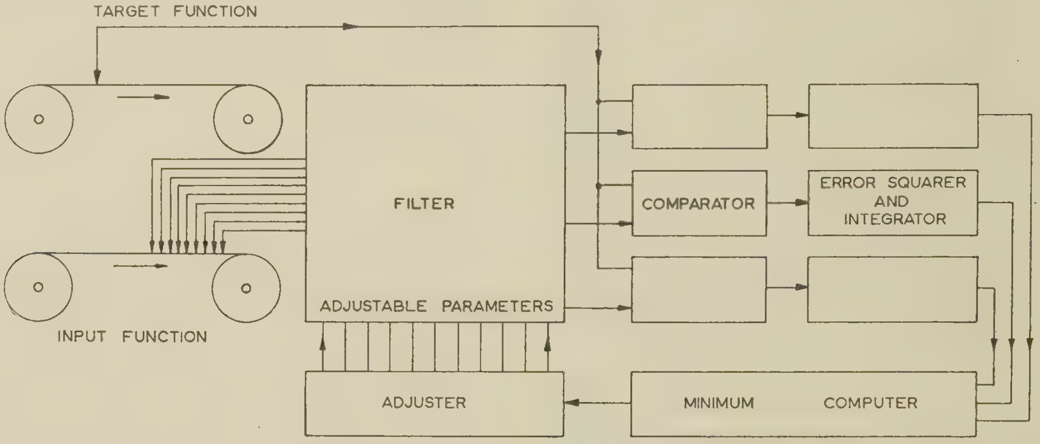


Fig. 2.—Block diagram of the learning filter and its training apparatus.

Fig. 2 is a block diagram of the filter with its teaching apparatus. The delay unit is a tape recorder with a number of staggered heads. The arithmetical units are contained in the filter block, and the adjustable parameters are represented by a number of shafts leading to an adjusting unit.

The tape also carries a track with the target function g(t). In the special case of prediction, this is read out by an advanced head from an input track. The output of the filter, together with the target function, is introduced into a comparator which computes the success criterion. The best success or error criterion has been an object of much discussion in the literature, but from a practical point of view there is only one choice, the criterion of least mean squares, as used by Kolmogoroff and by Wiener in their linear filter theories. That is to say we postulate

$$\overline{\{\mathrm{O}[\mathrm{f}(t)] - \mathrm{g}(t)\}^2} = \text{minimum} \quad . \quad . \quad . \quad (5)$$

The averaging has to be carried out over the ensemble available, i.e. over the length of the input and the target record.

The least-mean-squares criterion has the inestimable advantage that it gives a unique solution. The expression at the left of eqn. (5) is a positive definite quadratic form of the responses r; hence a solution will always exist and as the equations for the r's are linear there will be one solution only. These equations are given in full in Section 9.2.

The mean-square error as a function of the r's is a multidimensional elliptical paraboloid, and hence any rule by which this quantity is steadily diminished from run to run of the teaching apparatus must ultimately lead to the minimum. This is illustrated in Fig. 3 for the case of two variables. In this case the strategy of descent is obvious; the two variables must be adjusted alternately. In the general case there are many choices, and this problem will be discussed later.

The mean-square criterion, applied to a linear function of the r's, also has the advantage that the mean square error as a function of any one of the r's, with the others constant, is always a parabola and is completely determined by three points. To save time in our device the error comparators, squarers and integrators have been triplicated, so that in every run the integrated error-

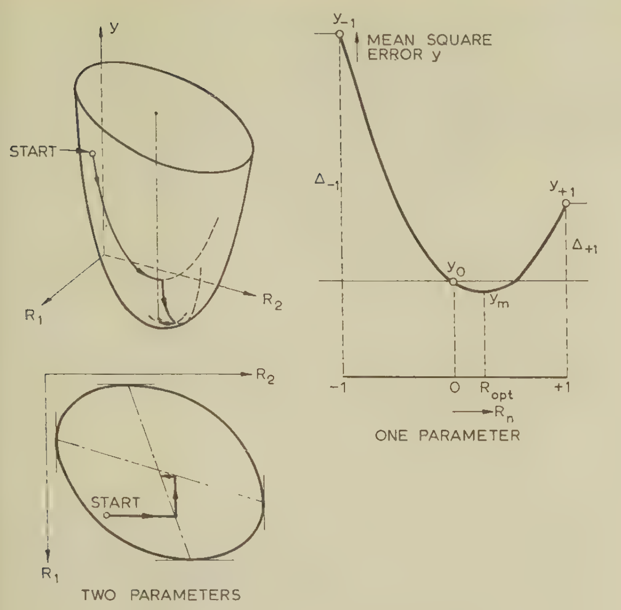


Fig. 3.—Descent to minimum error.

square is determined simultaneously for three values of the selected parameter r_n, for $r_n = 0$, and for the largest positive and negative values of it that are provided in the machine. From these values a minimum computer (Fig. 2) automatically determines the optimum value to which r_n must be set, and also the value of the minimum. The adjuster automatically sets r_n at this value and the next teaching run can start. As a general rule, if there are M adjustable parameters the number of runs required for approaching the optimum to within a small fraction will be of the order $\frac{1}{2}M^2$.

While it is clear that the adjustable parameters of a learning filter must figure in the output in a linear way, the choice of their co-factors as products of the samples is not necessarily the best. It is certainly the most economical at the present stage because multipliers are the most conveniently realized non-linear elements, and even these must be used sparingly. In a machine very much larger than the present one it might be advantageous to combine these into a complete set of orthogonal polynomials for a certain domain of the variables f_n, as Wiener[4] has proposed. It would, of course, be most uneconomical to take a separate set of multipliers for every term in every one of the orthogonal polynomials, but this could be avoided by attaching to each term a potentiometer with as many taps as there are polynomials in which it occurs.

What encourages us in the expectation that, if the filter has learned to minimize the error on a teaching record, it will do equally well when presented with another sample of the same stochastic series? The first assumption is, of course, that the series is ergodic, i.e. it has the same statistical parameters in any long sample. In addition, it is also necessary that in the learning process the machine shall abstract all relevant statistical parameters and reproduce them. It is shown in Section 9.2 that a machine which produces an expression according to eqn. (4) with a sufficient number of terms will base its minimization process on all autocorrelation functions with the target function g(t) to the same order. It can be shown that the statistical properties of any stochastic process are completely described by its autocorrelation functions of order 1, 2, 3 . . . This is intuitively clear if one realizes that an autocorrelation function of the order k is a histogram of the occurrence of groups of $k + 1$ peaks in the series. The ordinary, first-order autocorrelation function indicates by its peaks the occurrence of diads with a certain spacing, a second-order correlation function is a record of regularly occurring triads, and so on. If the series in question is a Markoff chain of finite order, its regularities can be exhausted by a filter with a finite number of terms.

The mathematical problems arising from Markoff chains of high order are excessively complicated and even their description is by no means simple. A first-order autocorrelation function is the function of one variable, but a second-order autocorrelation function requires a two-entry table, and so on. Our machine short-circuits the complicated description of the input process by high-order autocorrelation functions, and the description of its relation to the target function by cross-correlation functions, by going directly for the practical result: the optimum operator.

It has been mentioned that with M terms this will require a great number of repeated runs through the training samples, of the order $\frac{1}{2}M^2$, which is 5 000 in the somewhat unlikely case in which all 100 terms are of importance. It may well be asked whether the process could not be shortened by extracting the statistical parameters from the sample, which in principle could be done in one run, and feeding these into a digital computer which will solve the 100 simultaneous linear equations for the optimum coefficients.* The answer is that this could be done, but at the cost of a very large expenditure in electronic gear. By making use of the new multipliers developed for our machine, one can measure any autocorrelation coefficients by multiplying a set of samples and averaging the product by means of an integrator. Similarly, one can determine the cross-correlation coefficients by multiplying the product of samples with the target value and averaging [cf. Section 9.2, eqns. (24) and (25)]. The difficulty is only that in order to build up the equations for M terms one requires $\frac{1}{2}M(M + 1)$ values of the autocorrelation functions and M values of the cross-correlation functions. As it appears utterly impractical to provide such a number of multipliers and integrators, one could think of using the present number of multipliers, adding M integrators, and obtaining the coefficients in about $\frac{1}{2}M$ runs. Apart from requiring a considerable increase in the electronic gear, the objection to this is that the multipliers would have to be re-grouped from run to run and re-tested, because mistakes are likely to occur and errors can arise whenever complicated electronic equipment is switched round. For these reasons we consider our process of self-optimization to be the most economical at the present stage. It also has the advantage that it produces successive approximations, thereby giving a good idea of how many terms it is worth using in a filter in order to obtain a useful practical result.

(3) GENERAL DESIGN CONSIDERATIONS

(3.1) Requirements

The basic requirements for making a learning filter are:

(a) A storage medium containing f(t) and g(t) from which the two time series can be read out many times during the training period.

(b) An arrangement whereby delayed versions of the waveform f(t) can be obtained (f_1, f_2, f_3, etc.).

(c) Apparatus for multiplying wave samples with one another to form terms such as $f_{n1}f_{n2}$, $f_{n1}f_{n2}f_{n3}$, etc.

(d) Apparatus for multiplying the quantities f_{n1}, $f_{n1}f_{n2}$, etc., by variable coefficients, r_n, to form terms such as r_nf_n, etc.

(e) Apparatus for adding up the many terms of the operator polynomial, eqn. (4).

(f) Apparatus which computes the integrated square error y over one run of the stored record for three values of a given response coefficient.

* We are obliged to Dr. M. V. Wilkes for informing us that the large digital computer of the Cambridge Mathematical Laboratory requires only 7 minutes for solving 100 simultaneous linear equations.

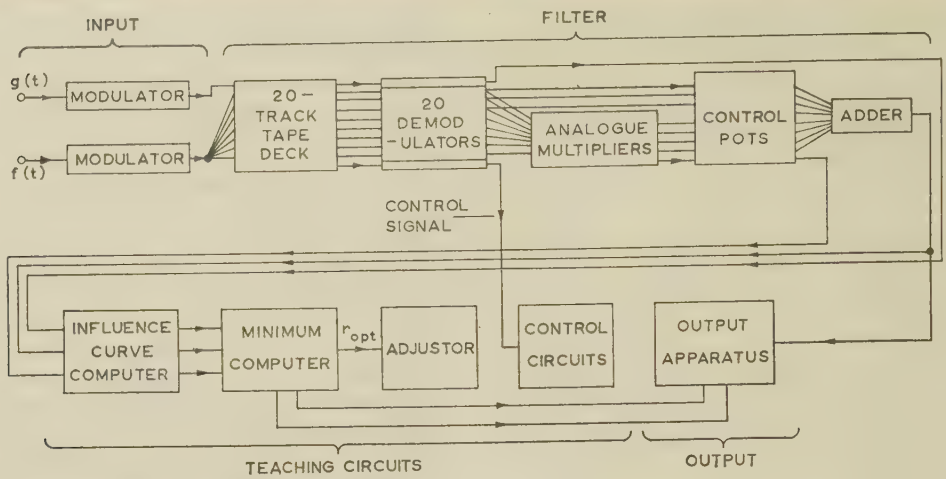


Fig. 4.—Detailed block diagram of the learning filter and training apparatus.

These are summarized in the block diagram, Fig. 2. If, in addition, the optimization is to be fully automatic with no human intervention, we must also have:

(*g*) A minimum computer which, from the three values of y obtained at the end of a run, computes the setting of a selected parameter, r_{opt}, and the minimum error, y_{min}.

(*h*) Apparatus which automatically adjusts the coefficients to the calculated values of r_{opt}.

(*i*) Sequence control of the whole process which arranges read-out of f(t) and g(t) at the correct time and initiates computation and adjustment.

(3.2) Realization of the Filter

It was clear from the first that the best storage medium would be magnetic tape. With normal amplitude recording, wear of the tape and 'dropouts' would give errors much greater than 1%, but at the standard speed of $7\frac{1}{2}$ in/sec and using a maximum signal frequency of 200 c/s, it would be possible to use some type of frequency modulation. Pulse rate modulation was finally chosen.

The necessary delayed versions of the signal are produced by recording the signal simultaneously on 18 tracks on the tape and using staggered pick-up heads for playback. A further track is needed for the teaching signal and a spare track for control purposes, making 20 in all.

It was originally considered that the Hall effect multipliers described by Kuhrt and Hartel[15] would prove suitable for generating products of the delayed signals. Extensive development was done by Dr. R. Woodcock, who produced several successful units, but the Hall effect multiplier was eventually discarded in favour of a new analogue multiplier invented by Dr. W. P. L. Wilby, which turned out to be cheaper and better in this application.

The obvious choice for the realization of the 100 variable coefficients was 100 potentiometers. By applying a signal to one end of a potentiometer and the same signal inverted to the other end, any coefficient between +1 and −1 can be obtained as the slider is moved from one end to the other. The 100 potentiometers must be servo-controlled in a fully automatic machine.

For adding up the many terms of the operator polynomial a conventional high-gain analogue adding amplifier was considered simplest.

The computation of the error y is a straightforward problem. At first it was thought that digital circuits were necessary to obtain the required accuracy. It was later realized that a simple modification would enable analogue circuits to do the job. This also changed the minimum computer from a digital to an analogue device.

Automatic sequence control of the whole process was realized by using relay and uniselector switching circuits.

The above account is summarized in the block diagram of the complete machine, Fig. 4.

(4) STRATEGY

It was pointed out in Section 2 that the learning process converges to a unique solution with all coefficients set to definite values, in whatever order the adjustment of the coefficients to their respective optima is carried out. With so many possible routes to the correct answer, we are still left with the task of deciding the best strategy of descent.

Strategy I.—A first possibility is to follow the Southwell relaxation rule,[16] always adjusting that coefficient which gives the greatest reduction in the mean-square error. Suppose we have 100 coefficients. We must first obtain the values of y_{min} for every coefficient and store them, together with the corresponding values of r_{opt}. This requires 100 runs of the tape record and consumes quite a long time. If the record is 100 sec long it takes almost 3 hours. We also need a store for 200 numbers, each known to at least 0·1%. Once we have the 100 values of y_{min} we must search through them to find the smallest. We must then set the corresponding coefficient to the appropriate value of r_{opt}. This is a straightforward computing problem and equipment was designed which would do the job. It became clear, however, that it would be very costly. Moreover, since each coefficient affects x_{opt} for every other coefficient, when we have set one r_{opt} the other 99 stored values are, strictly speaking, incorrect. Another 100 runs must therefore be performed before setting another coefficient. At this rate there would be about one adjustment every 3 hours and the machine might be worn out before ever reaching a solution. Time might be gained by assuming that, say, the 10 most important coefficients can be set using the stored values, and by assuming that the order of importance does not change after they are set. One would then only rank the coefficients in order of importance infrequently.

Strategy II.—Realization of the above difficulties led to a much simpler method. The really significant fact is that, although the machine takes a long time (100 sec) to obtain y_{min} and r_{opt} for a particular coefficient, it takes only a short time (1 sec) to adjust the coefficient. At the end of each run in Strategy I we obtained r_{opt} for one particular coefficient. Since it takes only 1 sec to adjust the coefficient we may as well do so. At the end of 100 runs we have now made 100 adjustments, including the coefficient which would have been adjusted by Strategy I. Furthermore, at each adjustment the effect of all

previous adjustments is taken into account, and so each step brings us safely towards the solution. It is clear that Strategy II must converge far more quickly than Strategy I. Also, we have eliminated the storage for 200 numbers and the apparatus required for ranking the y_{min}'s in order of importance—a very large item in the original plans.

Mechanization of the Learning Process.—Strategy II is readily mechanized, since the machine can deal blindly with one coefficient after another. However, almost certainly some coefficients will have only a small effect on the mean-square error and it is a waste of time to keep on adjusting them just as often as the more important ones. If we are to endow the machine with some discrimination we must do it by a method which requires little extra equipment and computing time. One method is to provide a single store which remembers the previous y_{min}. At the nth run we can then form

$$S = (y_{m_{n-1}} - y_{m_n})/y_{m_n}$$

as a measure of the significance of a given coefficient. If S is less than a certain pre-set amount we can instruct the machine to miss the coefficient on the next N cycles of the learning process. This is an instruction which is easily implemented. The machine can therefore be fully automatic, but even so it is desirable that manual control should be possible to enable the effect of different strategies and measures of significance to be tried out in practice without changing the design of the machine. This was kept in mind during the design of the control circuits.

(5) CONSTRUCTION OF THE FILTER

(5.1) Tape Recorder

The design of the recorder used for the magnetic tape records was based on a standard 16-track tape recorder. The deck accommodates 1 200 ft reels of tape and has two speeds, 7½ in/sec and 15 in/sec. The tape width was increased from 1 to 1¼ in to allow for 20 tracks. The recorder includes a control unit for manual control of play, record, spool and rewind. All the operations are controlled by relays and solenoids, thus making the deck suitable for inclusion in an automatic system.

(5.2) Delay Unit

The standard stack of 16 record/replay magnetic heads was replaced by the specially designed delay unit (Fig. 5). The essential part of the design is a cylindrical stack of 20 flat discs. A record/replay head is inset in each disc with a circular Mumetal magnetic screen above and below. A locating hole is provided in each disc so that the 20 heads can be lined up at the front of the cylinder. The tape runs around the front of the cylinder over a sector of 70°. At the back of each disc is a slot A. Two vernier scales B, each with a finger C which fits the slots A slide in horizontal guides which can be moved up and down vertically. Signal delays of up to 0·2 sec are obtained by recording with the heads in line and by staggering them on playback. Stagger is obtained by locking a disc in position using one vernier slider. The disc above is then rotated to the required position using the other vernier slider. This is then locked and the first vernier is removed and used to adjust the third disc, and so on. After all adjustments have been made the stack is clamped down using the lock-nut D.

(5.3) Pulse-Rate Modulation Record Playback System

The pulse-rate modulation system was designed around the variable-slope sawtooth oscillator described by Erath and Smith.[17] The circuit uses a pentode to discharge a capacitor. The voltage across the capacitor falls linearly until a trigger

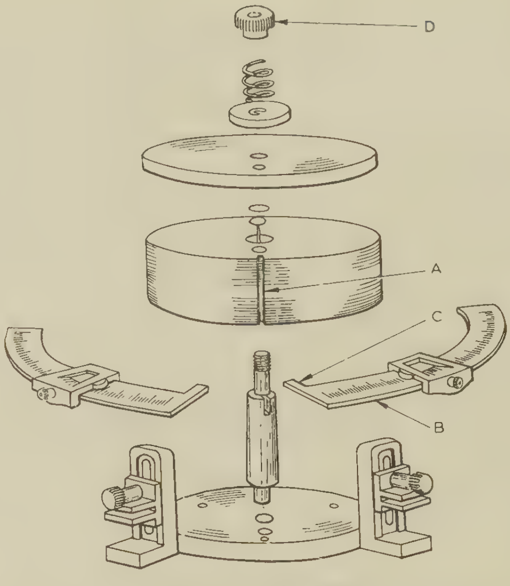


Fig. 5.—Delay unit of the magnetic tape recorder.

20 magnetic heads are set in metal discs, $\frac{1}{16}$ in thick, which can be rotated around a central pivot. The delays are adjusted by means of the two segments, each with one tooth engaging into a slot in a disc, whose angular position is read on verniers. The pile of discs is built up from the bottom to the top, and finally fixed by means of a locknut.

circuit operates, rapidly recharging the capacitor to its initial state. The rate of discharge varies according to the grid bias on the pentode. There is a linear relationship between pentode bias and repetition frequency over the range 1–7 kc/s. This range is compatible with a tape speed of 7½ in/sec. A current square wave is generated from the sawtooth wave and passed through the recording heads. The tape is saturated by the positive and negative current peaks.

On playback the signal from each head is passed via a 100 : 1 step-up transformer to a preamplifier. The output of the preamplifier is a series of short positive pulses corresponding to the rising edges of the original sawtooth wave. These pulses are fed to a pulse-rate demodulator. The first stage is a monostable pulse former which is triggered by the input pulses and produces a series of pulses of constant height and width, but occurring at a rate which depends on the recorded signal. The standard pulses are then integrated by a diode-pump pulse integrator[18] and the ripple which occurs at the pulse frequency is removed by an active RC filter. The filter response is 2% down at 200 c/s and 46 dB down at 1 kc/s. The main technical features are:

Recording current	= ± 50 mA
Recording frequency (square waves)	= 1–7 kc/s
Deviation ratio	= 75%
Tape speed	= 7½ in/sec
Total wow and flutter	= 0·3%
Total harmonic distortion ..	= − 40 dB
Signal amplitude	= 5 volts
Overall gain	= 1
Signal frequency band	= 5–250 c/s

(5.4) Servo-Controlled Potentiometers

Ninety-six precision potentiometers* have been installed for setting up the coefficients r_n, r_{n1n2}, etc. The 'in' spindles project downwards into an oil bath and each potentiometer has associated with it two magnetic clutches. Depending on which clutch is energized the slider can be rotated clockwise or anticlockwise. The motion is derived through gearing from a continuously rotating shaft.

* On loan from the Admiralty.

During each run equal and opposite signals corresponding to one term in eqn. (4) are applied to the two ends of a given potentiometer from a low-output-impedance phase splitter. Each slider is connected to an input of the analogue adding amplifier. The potentiometer whose influence curve is being computed contributes nothing to the sum, the contribution from its wiper being removed by an earthing contact. At the end of the run when it is required to set this potentiometer to the computed value of r_{opt}, a relay operates which disconnects its two ends from the phase splitter, and connects them to a ± 100-volt reference direct supply. The slider is disconnected from the adder and connected to one of the inputs of the potentiometer adjuster. The other input is fed with the voltage r_{opt}. A difference amplifier at the adjuster input compares the voltage at the slider with r_{opt} and the difference is amplified by a magnetic amplifier. The magnetic amplifier energizes one or other of the pair of magnetic clutches so that the potentiometer slider rotates in the direction which minimizes the difference between the two voltages. Using this servo-system it is possible to set up coefficients on the potentiometers with an error less than $0{\cdot}2\%$ of full scale.

(5.5) The Non-Linear Part of the Filter

(5.5.1) Multiplication Scheme.

As mentioned before, the number of terms tends to increase so steeply with the order that only non-linear processes with a strong 'perspective' can be realized with a reasonable number of terms.

A convenient method of limiting the terms is as follows. The first-order terms are weighted in inverse proportion to their delays and the significance of the $(N + 2)$th term is made zero. That is to say that the significance of the $(n + 1)$th sample, s_n, is

$$u_n = q\left(1 - \frac{n}{N+1}\right)$$

where q is a coefficient less than unity. The rule for the product terms is that they are cut off if their significance is less than that of the Nth sample. In order to retain the kth power of f_1, $q^{k-1} = 1/(N + 1)$.

Fig. 6 shows an arrangement of terms using 18 signal samples and using the full capacity of the filter. The samples 10 to 17 have been retained in the first-order terms with a view to fairly common problems in which non-linear effects are less important, while linear responses may 'ring' for a long time.

(5.5.2) Hall Multipliers.

Our original plan for the non-linear part of the filter was based on analogue multipliers of the Hall type, in particular of the semiconductor indium arsenide which has great advantages due to its low temperature sensitivity ($0{\cdot}1\%$ per deg C: Kuhrt and Hartel[15]; also Chasmar[19]). Moreover, it appeared at that time that the high efficiency attainable by the high conductance of indium arsenide would make it possible to drive one multiplier, without an intermediate amplifier, from another. This, however, had to be abandoned because the efficiency becomes sufficiently good for this only at such high powers that the heating of the Hall element becomes a problem.

The principle of the Hall multiplier is shown in Fig. 7(*a*). The unit developed by Dr. R. Woodcock, which used three double triodes, gave with inputs of ± 50 volts an output of ± 50 volts with errors of less than 1% of full output in the range 10–200 c/s.

The main advance in this design was the use of a new type of compensation circuit which is simpler than normal methods.

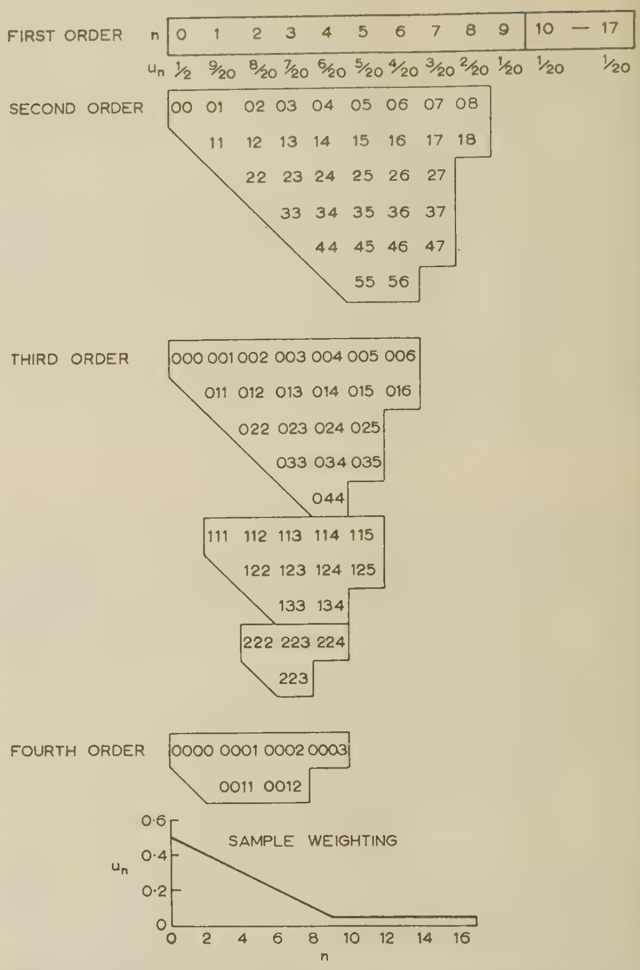


Fig. 6.—Scheme of allocating potentiometers and multipliers to a total of 94 terms of first, second, third and fourth order.

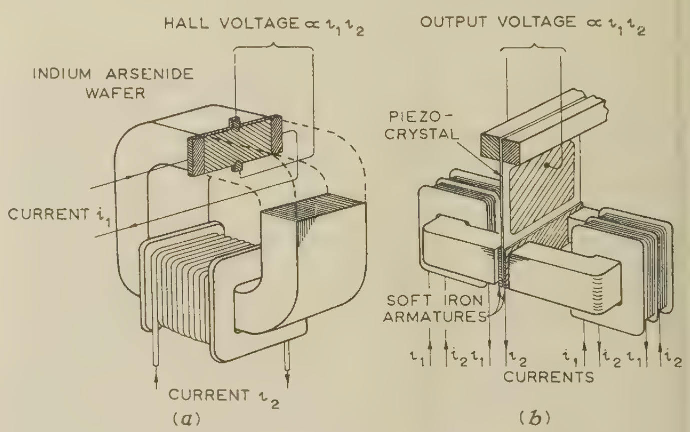


Fig. 7.—Analogue multipliers.

(*a*) Hall multiplier. (*b*) Piezo-magnetic multiplier.

This circuit is used to nullify the effect of field pick-up in the output circuit (or in the wafer current drive circuit if necessary). The usual method of reducing this effect in the output circuit is to bring the back pick-off contact lead through the air-gap directly over the other lead,[15, 20] or to have a compensation winding on the field core.[21] In the new method of compensation the back pick-off lead was brought back in two symmetrical

branches outside the air-gap to the two ends of a resistance wire which was tapped at a point at which the field pick-up was zero. This made it possible to save the extra air-gap, which is taken up in the usual design by the central lead and its insulation.

(5.5.3) **Piezo-Magnetic Multipliers.**

The operation of this new type of multiplier, invented by Dr. W. P. L. Wilby, is shown in Fig. 7(*b*). Four identical coils are placed on the limbs of two small silicon-iron C-cores. Opposite each C-core and separated from it by an air-gap is a rectangular silicon-iron armature. The armatures are attached to the free end of a bender-type piezo-electric crystal. The signals to be multiplied are represented by currents i_1 and i_2. In one magnetic circuit the coils are connected in the same sense and the magnetic field produced is proportional to $i_1 + i_2$. In the other magnetic circuit the coils oppose and the magnetic field is proportional to $i_1 - i_2$. The corresponding magnetic forces exerted on the armatures are proportional to $(i_1 + i_2)^2$ and $(i_1 - i_2)^2$. The magnetic forces are in opposite directions and subtract to give a net force on the piezo-crystal proportional to $(i_1 + i_2)^2 - (i_1 - i_2)^2$. The resulting output voltage from the piezo-crystal is given by $V_0 = K[(i_1 + i_2)^2 - (i_1 - i_2)^2] = K'i_1i_2$ and is a direct measure of the product.

To ensure constancy of K' the movement of the armatures must be a very small fraction of the air-gap length. The crystal is so stiff that this is readily ensured; its sensitivity is 5 volts per micron deflection at the tip. The reluctance of the air-gaps is made high compared with that of the iron path to minimize errors due to saturation. The maximum flux density in the armatures is 1 000 gauss. Accuracy of the product is ensured if the four coils are identical and if the effective air-gaps are balanced. Each coil has a resistance of 100 ohms and the multiplier input inductance is 70 mH. Each C-core with its two coils is cast into a semi-cylindrical epoxy-resin block, the pole faces being flush with the rectangular face of the block. The crystal, a corner-loaded square bimorph of Rochelle salt, with its armature assembly, is then mounted on three spacers on one of the blocks. The spacers define the air-gap between one armature and its adjacent pole faces. The air-gap of the other magnetic circuit is set up during assembly. The four coils are connected in series and an alternating current is passed through them, the output from the crystal being observed on an oscillograph. When the effective air-gaps are balanced the magnetic forces are equal and the crystal output is zero. The second block is adjusted until zero output (less than $\pm 0{\cdot}5\%$) is obtained and it is cemented in place using epoxy resin. This operation has proved quite straightforward in practice and has the great advantage that it makes unnecessary the provision of external zero adjusting arrangements such as a pre-set resistor in series with each coil.

The prototype multiplier measures 1 in diameter by $1\frac{1}{4}$ in long. It is mounted on an octal plug and sealed in a cylindrical dust cover. With a current input of 5 mA peak, an output of 250 mV peak is obtained. The resonant frequency of the crystal-armature assembly is 2 kc/s and this limits the high-frequency response of the unit. The peak in the frequency response due to resonance is removed by an *RC* network in the output amplifier. The response is shown in Fig. 8, and Table 1 gives a comparison of this multiplier with a Hall effect multiplier.

The modest input requirements and high output voltage of the multiplier make the design of the associated circuits a simple matter. A pair of cathode-followers (one double triode) provide a high input impedance and sufficient current to drive the coils. There is a 12-kilohm resistor in series with each multiplier input to allow a 100-volt swing at the cathode of the 12AT7 driver valve. A 12AX7 double triode is used as a negative-feedback

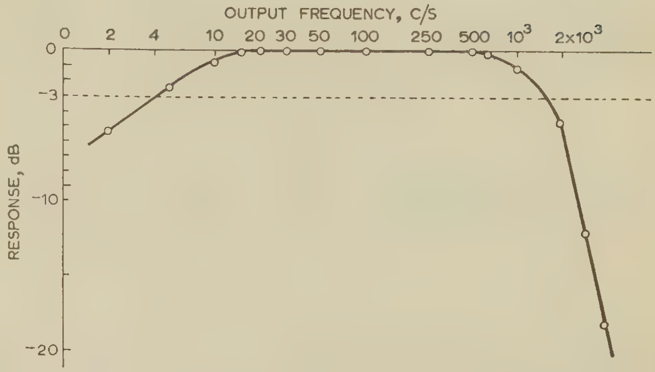


Fig. 8.—Overall frequency response of a piezo-magnetic multiplier, equalized by an *RC* filter.

Table 1

Comparison of a Hall Effect and a Piezo-Magnetic Multiplier with Maximum Input

Hall effect		Piezo-magnetic	
Field inductance	= 1 H	Input (1) inductance	= 70 mH
Field resistance	= 150 Ω	Input (1) resistance	= 200 Ω
Field current	= 40 mA	Coil current	= 5 mA
Wafer resistance	= 4 Ω	Input (2) inductance	= 70 mH
Wafer current	= 200 mA	Input (2) resistance	= 200 Ω
Output resistance	= 3·5 Ω	Coil current	= 5 mA
Output voltage	= 150 mV	Output voltage	= 250 mV

amplifier to amplify the output up to 50 volts peak. The output impedance is sufficiently low for the unit to feed up to four other multipliers, allowing a very flexible interconnection scheme. Two such units are accommodated on a single chassis measuring $19 \times 1\frac{3}{4} \times 3$ in. The main details of the complete multiplier unit are:

Peak input	= ± 50 volts
Peak output	= ± 50 volts
Maximum output frequency ..	= 1 kc/s
Overall total harmonic distortion at peak output	= − 46 dB
Long-term amplitude drift over 1 month under normal laboratory conditions	= less than 2%
Stabilized power supplies	= ± 300 volts, 12 mA

An experimental unit was also made using 6 OC71 transistors. Input and output voltages were 1 volt peak and the total power consumption was only 100 mW. It was found to be less stable and more expensive than the valve circuit.

(5.6) Teaching Circuits

(5.6.1.) **Influence Curve Computer and Minimum Computer.**

The teaching circuits can be divided into two parts, the influence curve computer, and the minimum computer.

The influence curve computer consists of the comparator, squarers and integrators. The quantities computed are stored as voltages on the integrators. At the end of each run these voltages are fed into the minimum computer which computes the value of r_{opt} and y_{min}. The value of r_{opt} is fed into the adjusting unit which sets the servo-controlled potentiometer to this value.

The influence curve is parabolic and it can therefore be completely specified by three points. The three points taken

have the abscissae $r = -1$, 0, $+1$, and the corresponding ordinates y_{-1}, y_0 and y_{+1}. From these one obtains

$$r_{opt} = \frac{1}{2}\frac{y_{-1} - y_{+1}}{y_{-1} + y_{+1} - 2y_0}$$

It is seen that very accurate computation would be necessary if r_{opt} were computed directly from y_{-1}, y_0 and y_{+1}. However, in terms of the differences $\Delta_{-1} = y_{-1} - y_0$ and $\Delta_{+1} = y_{+1} - y_0$,

$$r_{opt} = \frac{1}{2}\frac{\Delta_{-1} - \Delta_{+1}}{\Delta_{-1} + \Delta_{+1}}$$

and $$y_{min} = y_0 - \tfrac{1}{2}r_{opt}\tfrac{1}{2}(\Delta_{-1} - \Delta_{+1})$$

The values $\frac{1}{2}(\Delta_{-1} - \Delta_{+1})$, y_0 and $\Delta_{-1} + \Delta_{+1}$ could be formed in the influence curve computer and stored on the integrators.

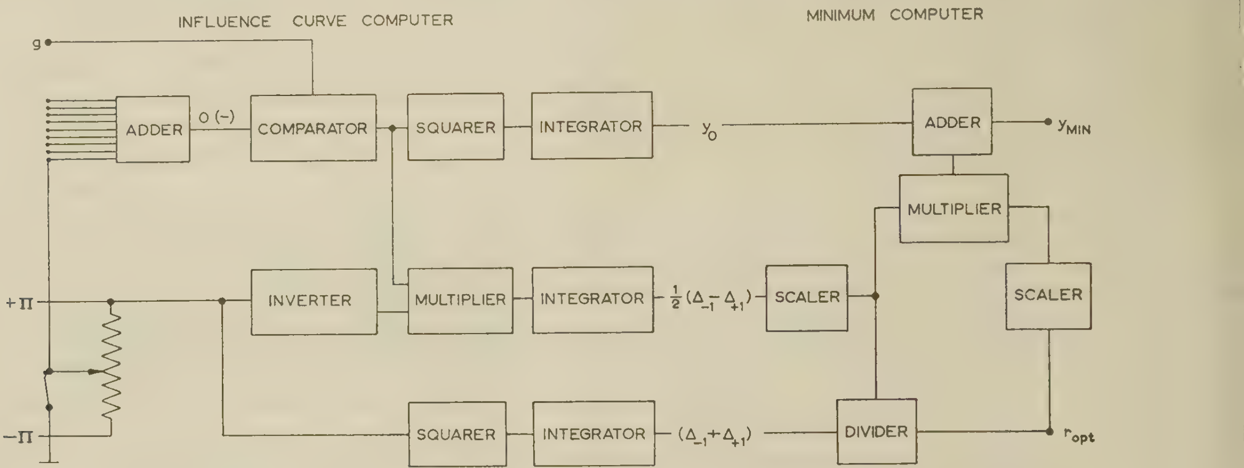


Fig. 9.—The arithmetical circuits for computing the optimum setting of any parameter, r, and the value of the minimum.

It will now be shown that these values are simple functions of terms in the polynomial, eqn. (4).

Let O(−) be the output of the filter with one coefficient r set to zero, i.e. one term in eqn. (4) suppressed. Let Π be that term with the coefficient r set to +1, i.e. $\Pi = f_{n1} \ldots f_{nk}$, and let g be the target signal. Then

$$y_{+1} = \overline{[O(-) + \Pi - g]^2} \quad y_0 = \overline{[O(-) - g]^2}$$

$$y_{-1} = \overline{[O(-) - \Pi - g]^2}$$

A little calculation gives

$$\Delta_{-1} + \Delta_{+1} = y_{-1} + y_{+1} - 2y_0 = \overline{2\Pi^2}$$

$$\tfrac{1}{2}(\Delta_{-1} - \Delta_{+1}) = \tfrac{1}{2}[y_{-1} - y_0 - (y_{+1} - y_0)]$$
$$= -\overline{2\Pi[O(-) - g]}$$

This results in the block diagrams for the influence curve computer and the minimum computer shown in Fig. 9, which are greatly simplified by making use of the above relations.

The divider is a high-gain d.c. amplifier with a multiplier in the feedback loop. With sufficiently high gain the transfer function of a feedback amplifier is the reciprocal of the feedback factor. With this type of divider the input to the multiplier (the denominator) must always be of one sign, i.e. positive. Reversal of the sign would reverse the phase of the feedback making it positive and the divider would oscillate. The denominator must not become small for as it tends to zero the output tends to infinity. But neither of these factors limits its use in this arrangement since $\Delta_{-1} + \Delta_{+1} = \overline{2\Pi^2}$ which must always be positive and does not change as the optimization proceeds. The divider will not saturate if r_{opt} lies between −1 and +1. If it lies outside this range a warning light indicates that the target track signal must be re-scaled.

(5.6.2) **Triangular Wave Multiplier.**

An analogue multiplier was required for use in the influence curve computer and the minimum computer. This had to have a flat response from 0 to 200 c/s. The method chosen was a 'quarter squares' multiplier using a gated triangular wave:

$$xy = \tfrac{1}{4}(x + y)^2 - (x - y)^2$$

The basic circuit was that used by Meyer and Davies.[22] This had the advantage over many others that all gating was with reference to earth. A simplification was made to the circuit reducing the number of amplifiers by one. Originally two inversion operations were carried out before clipping and the clipped signals were then added. The clipping diodes can be reversed and the signals summed and inverted before clipping. This gives the modified block diagram, Fig. 10.

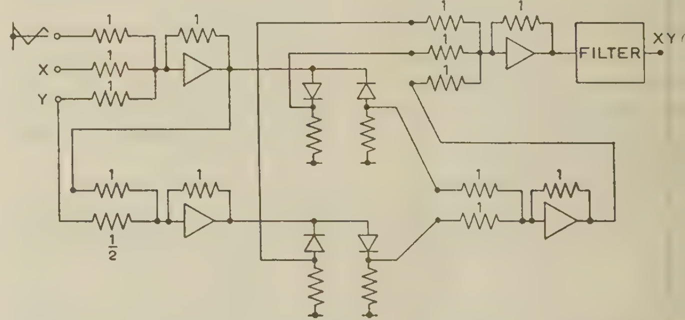


Fig. 10.—Triangular-wave multiplier as used in the arithmetical circuits.

The amplifier developed for the multiplier was a cascode stage directly coupled to a cathode-follower with feedback from the cathode-follower output to the input grid in the normal way. The frequency of the triangular wave was 5 kc/s and the multiplier gave for inputs of ±5 volts an output of ±5 volts with an error of less than 1% from 0 to 600 c/s.

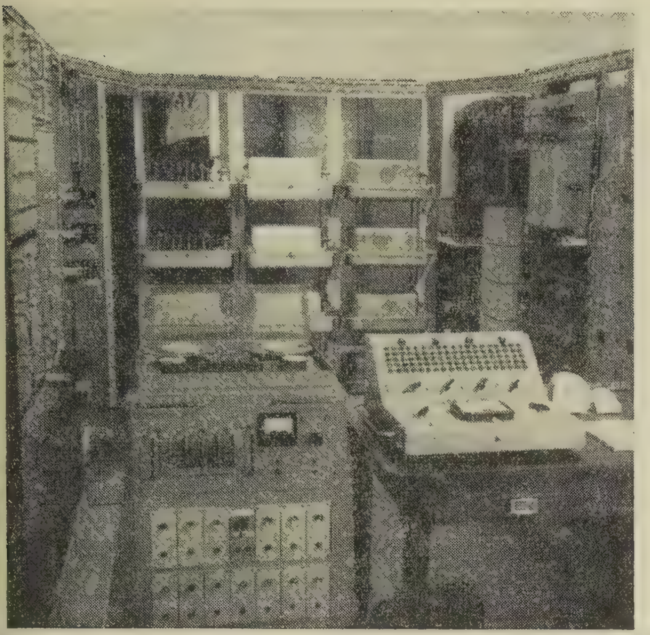


Fig. 11.—View of the learning filter, October, 1959.

In the left foreground is the 20-track tape recorder, at the right is the control panel. Around the racks from left to right: demodulators, power supply, multiplier racks, control of potentiometers, 96 servo-controlled potentiometers, selectors, and arithmetical circuits.

(5.7) Control Circuits

Control of the apparatus is effected from a control panel seen in Fig. 11, which shows a view of the whole device. The control panel has a neon indicator lamp for each coefficient. This lamp flashes to indicate the coefficient whose influence curve is being computed and glows steadily if a coefficient is being temporarily ignored in the learning sequence. This coefficient is selected by four 25-way uniselectors which step in series, selecting one of a hundred possible positions. Each coefficient has a 'miss' relay associated with it which, if energized, prevents the uniselectors from stopping at a coefficient which is to be ignored. Each coefficient has a coupling relay which disconnects its corresponding servo-controlled potentiometer from the filter and connects it to the adjuster. The circuits are made through the uniselectors, so that one adjuster suffices for controlling all the potentiometers.

The learning sequence can be started at any coefficient. All measurements and voltages in the machine are referred to a meter on the control panel, which reads the value of any selected coefficient and the output voltages of the influence curve and minimum computers. The operation can be fully automatic, in which case the coefficients which are not missed are adjusted in sequence at the end of consecutive runs during the rewind period. Provision is also made for part-automatic or fully manual control.

(5.8) Output Apparatus

The teaching track and the output of the learning filter can be viewed together on a double-beam oscillograph. The signals can also be recorded on film using an oscillograph camera.

An auxiliary twin-track 3-speed tape recorder is also available for recording the results. With some filtering problems it is interesting to listen to the input and output of the filter by passing them through an a.f. amplifier and loudspeaker. Before doing so it is preferable to use the auxiliary tape recorder to frequency-shift the signals into the normal audio band. The above methods are suitable for recording the final result and for observing the performance of the filter at a few intermediate stages of the learning process. A running check on progress is also desirable and this is obtained by means of a teleprinter, which at the end of each run prints out the run number, coefficient number, r_{opt} and y_{min}.

(5.9) Preliminary Results

The first problems designed to test the system as a whole were

(*a*) Transforming a sine wave to another, different in phase and amplitude.

(*b*) Filtering a sinusoidal signal with superimposed noise.

These are tests of the linear part only.

Phase-shifting and scaling a sinusoidal signal.—For this test only two filtering tracks were used. The signal on channel 2 was delayed by moving the pick-up head to give a lag of approximately 120°. The target track was delayed on channel 1 to give a lag of 60°. The filtering and teaching channel signal amplitudes were the same (5 volts peak). The correct solution is known to be -1, -1. The two channels were initially given coefficients of $+1$, as wrong as possible. The progress of the optimization is seen in Fig. 12. Although only two filtering channels were used, the filter did not optimize in only 2 runs and it was necessary to return again to the previous coefficient after the setting of one. This is the relaxation routine which will always be necessary in even the simpler cases. It is noted that the coefficients are not orthogonal. The coefficients finally set themselves to equal values as was expected, since the phase of the teaching track was halfway between the two filtering tracks.

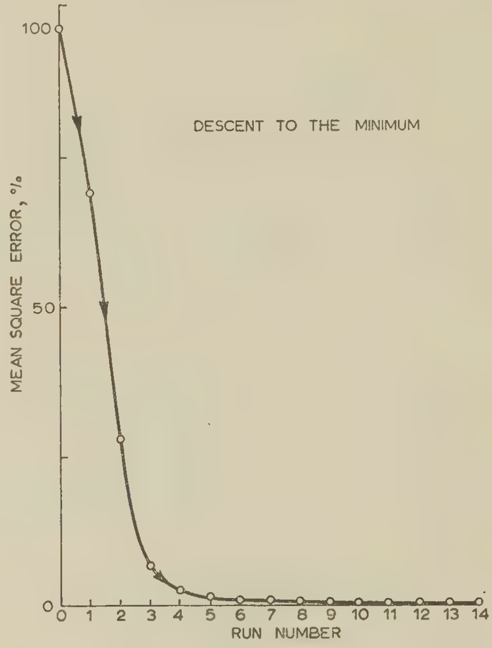


Fig. 12.—First test. Phase shifting and scaling of two sine waves to match a third sine wave.

Filtering a sine wave from sine wave plus noise.—A 200 c/s sine wave of 5 volts peak was recorded on the teaching track and at the same time was added to noise and recorded on six filtering tracks, the magnetic heads being in line for recording. The amplitude of the noise power, which was within the band 10–300 c/s, was adjusted to be equal to the sine-wave power and added to it. The power was measured by squaring the signal or the noise and noting the d.c. component.

The magnetic heads on the tape deck were then staggered with the teaching track, channel 1 was set with the same phase and channels 2–6 were given delays which were integers of $1/(2F)$, where $F = 200$ c/s. The machine was then allowed to optimize itself. The result is seen in Fig. 13.

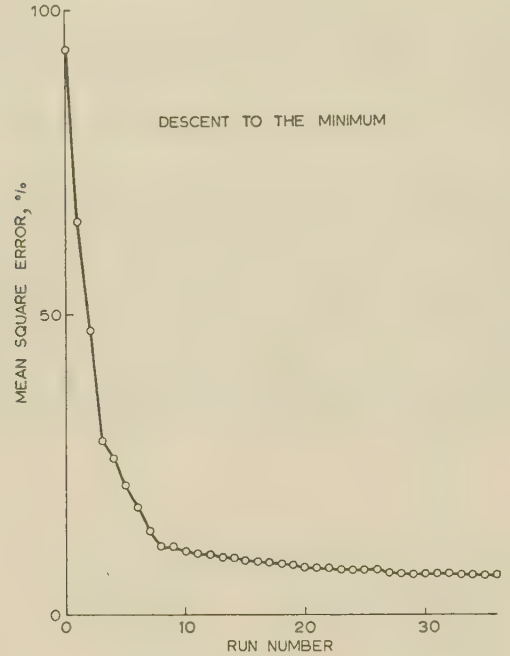


Fig. 13.—Second test. Filtering a noisy signal. Descent to the minimum error as a function of teaching runs.

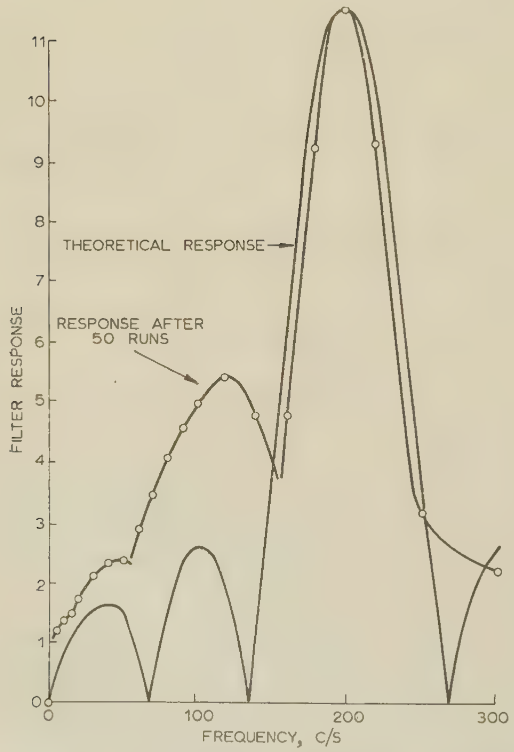


Fig. 14.—Second test. Filtering a noisy signal. The ideal filter response and its approximation achieved in 50 teaching runs.

It is well known that the optimum solution is with the coeffi cients equal and alternating in sign. With six terms the maxi mum possible improvement in the signal/noise power ratio i 6 times. (Addition of the six sinusoidal components due to th correlation between signal samples gives an increase in powe equal to 6^2. Addition to the noise from the six sampling point gives a noise power of 6 times only, since, if the noise is trul random, there is no correlation between the sampling poin signals.)

The frequency response of the filter (Fig. 14) was taken afte 50 runs and compared with the theoretical characteristic. Th graphs were compared for 200 c/s and below, as the top frequenc the filter could handle was 200 c/s and above this frequency ther was sharp cut-off in the demodulators of the tape playbac system. The response curve shows the main peak at 200 c/ and the subsidiary peaks at about 105 and 35 c/s. The zer are only weakly marked.

After this characteristic was taken the machine was allowe to run for another 180 runs, 230 in all. It was most interestin to observe that, while the mean-square error decreased only b about 0·5% in these additional runs, the frequency respons kept steadily approaching the theoretical shape. It is a mos gratifying proof of the reliability of the machine that it coul steadily improve itself on such a slight gradient of the mean square error, for such a long series of runs.

The relatively simple tests were made as the theoretical resu was known or could be calculated and it was then possible t check the filter's performance more readily.

(6) CONCLUSION AND OUTLOOK

After a long period of development and construction, ou learning filter has just started operating, and has so far bee tested only on simple problems. The further programme is

(*a*) To build it up to the full capacity of 96 terms,
(*b*) To construct input equipment enabling the machine to opera on statistical material contained in tables, graphs or non-standar tapes.
(*c*) To automatize the learning process completely.

It is to be expected that weaknesses in certain parts will revea themselves in long teaching runs, which will have to be elimi nated. After this it is intended to put the machine to use on great variety of engineering and scientific problems of increasin complication, testing its capabilities and its limitations, an studying the possibilities of even larger and more powerfu machines.

(7) ACKNOWLEDGMENTS

We were enabled to carry out this research and to construc the apparatus by the assistance of the Admiralty Surface Weapon Establishment, Portsdown, Hants. Drs. W. P. L. Wilby an R. Woodcock received Bursaries sponsored by an Admiralt Research Contract. Our thanks are due in particular to Mr. . Anderson, Chief Scientist, and to Mr. R. Benjamin, from whos unfailing interest and valuable advice this work has great benefited.

(8) REFERENCES

(1) Kolmogoroff, A.: 'Interpolation and Extrapolation o Stationary Series', *Bulletin de l'Académie des Sciences d l'U.R.S.S., Séries Mathématiques*, 1942, **2**, p. 3.
(2) Wiener, N.: 'The Extrapolation, Interpolation and Smooth ing of Stationary Time Series' (Wiley, 1949).
(3) Wiener, N.: 'Response of a Nonlinear Device to Noise M.I.T. Radiation Laboratory Report No. 129, 1942.
(4) Wiener, N.: 'Nonlinear Problems in Random Theor (Wiley, 1958).

(5) BOSE, A. G.: 'A Theory of Nonlinear Systems', M.I.T. Research Laboratory of Electronics Report No. 309, 1956.
(6) ZADEH, L. A.: 'Optimum Nonlinear Filters', *Journal of Applied Physics*, 1953, **24**, p. 396.
ZADEH, L. A., and RAGAZZINI, J. R.: 'An Extension of Wiener's Theory of Prediction', *ibid.*, 1950, **21**, p. 645.
'Optimum Filters for the Detection of Signals in Noise', *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1223.
(7) SINGLETON, H. E.: 'Nonlinear Systems with Sampled Input', M.I.T. Research Laboratory of Electronics Report No. 160, 1951.
(8) WHITE, W. D.: 'The Role of Nonlinear Filters in Electronic Systems', Proceedings of the National Electronics Conference, 1953, **9**, p. 505.
(9) GABOR, D.: 'Communication Theory and Cybernetics', *Transactions of the Institute of Radio Engineers*, 1954, **CT–1**, No. 4, p. 19.
(10) TUSTIN, A.: 'Mechanism of Economic Systems' (Heinemann, 1953).
(11) NYQUIST, H.: 'Certain Factors affecting Telegraph Speed', *Bell System Technical Journal*, **3**, p. 324.
KUEPFMUELLER, K.: 'Transients in Wave Filters', *Elektrische Nachrichten-Technik*, 1924, **1**, p. 141.
(12) GABOR, D.: 'Theory of Communication', *Journal I.E.E.*, 1946, **93**, Part III, p. 429.
(13) SHANNON, C. E.: 'Communication in the Presence of Noise', *Proceedings of the Institute of Radio Engineers*, 1949, **37**, p. 10.
(14) OSWALD, J.: 'Les signaux à spectra limités et leurs transformations', *Câbles et Transmission*, 1950, **4**, p. 215.
(15) KUHRT, F.: 'Properties of Hall Generators', *Siemens Zeitschrift*, 1954, **28**, p. 370.
HARTEL, W.: 'Application of Hall Generators', *ibid.*, p. 376.
(16) SOUTHWELL, R. V.: 'Relaxation Methods' (Oxford University Press, 1946).
(17) ERATH, L. W., and SMITH, F. C.: 'A Slope Modulator for FM Recording of Analog Data on Magnetic Tape', *Transactions of the Institute of Radio Engineers*, 1954, **PGRTRC–2**, p. 20.
(18) EARNSHAW, J. B.: 'The Diode Pump Integrator', *Electronic Engineering*, 1956, **28**, p. 26.
(19) CHASMAR, R. P., and COHEN, E.: 'An Electrical Multiplier utilizing the Hall Effect in Indium Arsenide', *ibid.*, 1958, p. 661.
(20) LOFGREN, L.: 'Analogue Multiplier Based on the Hall Effect', *Journal of Applied Physics*, 1958, **29**, p. 158.
(21) KEISER, G. L.: 'A Compact Multiplier puts the Hall Effect to Work', *Control Engineering*, November, 1955, **2**, p. 94.
(22) MEYERS, R. A., and DAVIS, H. B.: 'Triangular-Wave Analog Multiplier', *Electronics*, August, 1956, p. 182.

(9) APPENDICES

(9.1) Bandwidth-Limited Signals and their Processing

(9.1.1) The Whittaker–Shannon Sampling Theorem.

A bandwidth-limited time function is one whose Fourier spectrum is zero outside some frequency band of width F. For convenience we will consider, as usual, a real signal in a frequency band $\pm F$. One half of this contains redundant information. It has been shown[9] how this can be transformed into a complex 'analytical' signal extending over a waveband F only. (Passage from double sideband to single sideband.)

By Whittaker's interpolation formula,* introduced by Shannon[13] into communication theory, one can associate with any arbitrary time function $f(t)$ a cardinal function $[f(t)]$ whose Fourier transform is identical with that of $f(t)$ inside $\pm F$ but is zero outside. Whittaker's formula is

$$[f(t)] = \sum_{n=-\infty}^{\infty} f(t_n)u(t - t_n) \quad . \quad . \quad . \quad . \quad (6)$$

where the sampling points t_n are spaced by the Nyquist intervals $1/2F$ and u is the function

$$u(\tau) = \sin 2\pi F\tau / 2\pi F\tau \quad . \quad . \quad . \quad . \quad . \quad (7)$$

As the coefficients $f(t)_n$ are the samples of $f(t)$ at the points $t = t_n$, this has been called by Shannon the sampling theorem.

(9.1.2) The Expansion Theorem.

Though the sampling theorem is very useful for simple mathematical discussion, it suffers from the disadvantage that the operation

$$f(t) \rightarrow [f(t)]$$

is not physically realizable on a 'live' signal, of which we know only the past. The expansion theorem[9] goes one step nearer to physical reality.

Let $r(t)$ be the impulse response function of a linear filter to a Dirac pulse at $t = 0$. Following van der Pol and Bremmer we will call this the time response, and its Fourier transform, $R(f)$, the frequency response. For a physical filter, $r(t)$ must be zero for all negative times. The expansion theorem states that an observer beyond a filter whose frequency response is zero outside a certain interval, F, cannot tell the difference between a time function $f(t)$ and its cardinal function

$$[f(t')] = \sum_{-\infty}^{t_n = t} c_n r(t' - t_n) \quad t' \leqslant t \quad . \quad . \quad . \quad (8)$$

This is an expansion in terms of the response function r with expansion coefficients which are not in general samples of $f(t)$. They can, however, be interpreted as samples of a pre-emphasized signal. If the Fourier transform of $f(t)$ is $F(f)$ the transform of the pre-emphasized signal is $F(f)/R(f)$ inside the band $\pm F$ and zero outside it. Unlike eqn. (6), eqn. (8) is an expansion of a time function in terms of its past only, valid up to the present moment, t.

Proof.—The operation O of a linear filter with time response r on a time function $f(t)$ can be written

$$O\{f(t)\} = \int_{-\infty}^{t} f(\tau)r(t - \tau)d\tau \quad . \quad . \quad . \quad . \quad (9)$$

As the response r is zero for negative arguments, we could as well replace the upper limit t by any larger time. Similarly, we could replace the lower limit, $-\infty$, by any time earlier than the start of the signal f. We set limits $\pm\frac{1}{2}T$ in eqn. (9) and also in eqn. (8). Eventually we will extend T to infinity.

Replacing f in eqn. (9) by the expansion (8) we obtain

$$O\{[f(t)]\} = \int_{-\frac{1}{2}T}^{\frac{1}{2}T} \sum_{-\frac{1}{2}T}^{\frac{1}{2}T} c_n r(\tau - t)r(t - \tau)d\tau \quad . \quad . \quad (10)$$

Thus, the error owing to the substitution of $[f(t)]$ for $f(t)$ is

$$O\{[f(t)]\} - O\{f(t)\} = \int_{-\frac{1}{2}T}^{\frac{1}{2}T} \left[\sum_{-\frac{1}{2}T}^{\frac{1}{2}T} c_n r(\tau - t_n) - f(\tau) \right] r(t - \tau)d\tau \quad (11)$$

* WHITTAKER, E. T.: 'On the Functions which are Represented by the Expansions of the "Interpolation Theory"', University of Edinburgh, Mathematical Department Research Paper No. 8, 1915.

In the limit, for very large intervals T we can use the Fourier integral formulae

$$\int_{-\infty}^{\infty} \mathrm{f}(\tau)\mathrm{r}(t-\tau)d\tau = \int_{-\infty}^{\infty} \mathrm{F}(f)\mathrm{R}(f)\varepsilon^{2\pi jft}df$$

$$\int_{-\infty}^{\infty} \mathrm{r}(\tau - t_n)\mathrm{r}(t-\tau)d\tau = \int_{-\infty}^{\infty} \mathrm{R}^2\varepsilon^{2\pi jf(t-t_n)}d$$

Hence,

$$\mathrm{O}\{[\mathrm{f}(t)]\} - \mathrm{O}[\mathrm{f}(t)] = \int_{-\infty}^{\infty} \mathrm{R}(f)\varepsilon^{2\pi jft}\left[\sum_{-\frac{1}{2}T}^{\frac{1}{2}T} c_n \mathrm{R}(f)\varepsilon^{-2\pi jft_n} - \mathrm{F}(f)\right]df \quad . \quad . \quad . \quad . \quad (12)$$

The mean-square error, averaged over a time T, with which we now go to the limit ∞, becomes

$$\overline{[\mathrm{O}\{[\mathrm{f}(t)]\} - \mathrm{O}\{\mathrm{f}(t)\}]^2} =$$

$$\lim_{T\to\infty} \frac{1}{T}\int_{-\frac{1}{2}T}^{\frac{1}{2}T} dt \iint_{-\infty}^{\infty} \mathrm{R}(f_1)\mathrm{R}(f_2)[f_1][f_2]\varepsilon^{2\pi j(f_1-f_2)t}df_1 df_2$$

where we have introduced abbreviations for the square bracket expression under the integral sign in eqn. (12). The integration over t can be easily carried out and the resulting expression for the mean-square error is

$$\lim_{T\to\infty} \frac{1}{T}\int_{-\infty}^{\infty} |\mathrm{R}(f)|^4 \left|\sum_{-\frac{1}{2}T}^{\frac{1}{2}T} c_n \varepsilon^{\pi jnf/F} - \frac{\mathrm{F}(f)}{\mathrm{R}(f)}\right|^2 df \quad . \quad (13)$$

Assume now that the frequency response, $\mathrm{R}(f)$, of the filter vanishes outside $\pm F$. Inside the frequency band $\pm F$ we can make the second factor in the integrand vanish by making the coefficients c_n equal to the coefficients of the Fourier series expansion of $\mathrm{F}(f)/\mathrm{R}(f)$ in this interval. There are $2FT$ Fourier coefficients in this interval, as many as there are independent data in the frequency band F and in the time interval T. If $\mathrm{R}(f)$ = a constant in the range $\pm F$ the c_n become identical with the Nyquist samples, otherwise they are samples of the pre-emphasized signal. This proves the expansion theorem.

This theorem, however, suffers from the deficiency that the frequency response of a physical filter cannot be made exactly zero outside any finite frequency band.

Proof.—It is known that the real and imaginary parts of the frequency response of a realizable (physical) filter are Hilbert transforms of one another, i.e. they are connected by the reciprocal relations[12]

$$\mathscr{I}\mathrm{R}(f) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\mathscr{R}\mathrm{R}(\nu)}{f-\nu}d\nu \quad \mathscr{R}\mathrm{R}(f) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\mathscr{I}\mathrm{R}(\nu)}{f-\nu}d\nu \quad (14)$$

The integrals are understood to be taken along the indented axis to eliminate the pole at $f = \nu$ (Cauchy's principal value).

Assume now that the real part of $\mathrm{R}(f)$ is band-limited. Outside this band we can expand the denominator

$$\frac{1}{f-\nu} = \frac{1}{f}\left[1 + \frac{\nu}{f} + \left(\frac{\nu}{f}\right)^2 + \ldots\right]$$

and obtain

$$\mathscr{I}\mathrm{R}(f) = \frac{1}{\pi f}\left[\int_{-F}^{F} \mathscr{R}\mathrm{R}(\nu)d\nu + \frac{1}{f}\int_{-F}^{F} \mathscr{R}\mathrm{R}(\nu)\nu d\nu + \frac{1}{f^2}\int_{-F}^{F} \mathscr{R}\mathrm{R}(\nu)\nu^2 d\nu + \ldots\right] \quad (15)$$

The coefficients in this expansion are the moments of the rea part of the frequency response. If the imaginary part of th response is to vanish everywhere outside $\pm F$ all moments mus vanish, and it is known that this means that the integrand mus vanish everywhere.

This proves that with physical filters the expansion theorer is never exact but only an approximation. But in fact it ca be a very good approximation. If one allows a sufficient dela between the impulse and the maximum of the filter response, physical filter can approximate to a bandwidth-limited filter t any degree. In practice a delay of a few Nyquist intervals i sufficient, and this is almost always allowable, except in som problems in which rapid action is required (e.g. certain problem of short-time prediction and recognition, mainly of militar interest). We will therefore use the expansion theorem as useful approximation.*

(9.1.3) **Processing of Band-Limited Information.**

Processing Theorem.—Any single-valued operation on the pas of a band-limited time function is equivalent to an operation o Nyquist samples.

Proof.—It is known† that the most general single-valued time-invariant operation on the past of a time function, f(t which converts it into another time function, g(t), can be repre sented as a functional power series

$$\mathrm{g}(t) = \int_{-\infty}^{t} \mathrm{f}(\tau)\mathrm{h}_1(t-\tau)d\tau + \int_{-\infty}^{t}\int_{-\infty}^{t} \mathrm{f}(\tau_1)\mathrm{f}(\tau_2)\mathrm{h}_2(t-\tau_1, t-\tau_2)d\tau_1 d\tau_2 + \ldots \quad (16)$$

Causality (physical realizability) requires that all response func tions h_n must vanish if any one of their arguments is negative, i.e

$$\mathrm{h}_n(t-\tau_1, \ldots t-\tau_k \ldots) = 0 \text{ if any } \tau_k > t \quad (17)$$

Consequently, we could replace the upper limits t in all th integrals by $+\infty$ without any change in the value of g(t).

Let us interpose a filter with time response r, frequency ban $\pm F$ between f(t) and the general operator. By the expansio theorem (and therefore *cum grano salis*) we can replace f(t) b its cardinal function

$$[\mathrm{f}(t')] = \sum_{-\infty}^{t_n \leqslant t} c_n \mathrm{r}(t'-t_n) = \sum_{n=0}^{n=\infty} c_n \mathrm{r}(t'-t+n\tau_0) \qquad t' \leqslant t \quad (18)$$

The first expression is as before. In the second we have intro duced sliding sampling points at $t_n = t, t-\tau_0, \ldots t-n\tau_0 \ldots$ where τ_0 is the Nyquist interval and n increases towards the pas In other words the operator now operates on the signal sample delayed by $0, \tau_0 \ldots n\tau_0 \ldots$ Because of the 'physical' propert of the response r, we could replace the lower limit $n = 0$ b $n = -\infty$ without a change in the value of the cardinal functio

Substituting this into eqn. (16) we obtain for the first term

$$\int_{-\infty}^{t} \mathrm{f}(\tau)\mathrm{h}_1(t-\tau)d\tau = \int_{-\infty}^{\infty} \mathrm{f}(\tau)\mathrm{h}_1(t-\tau)d\tau$$

$$= \int_{-\infty}^{\infty} \sum_{0}^{\infty} c_n \mathrm{f}(\tau - t + n\tau_0)\mathrm{h}_1(t-\tau)d\tau$$

$$= \sum_{0}^{\infty} c_n \int_{-\infty}^{\infty} \mathrm{r}(t-\tau+n\tau_0)\mathrm{h}_1(t-\tau)d\tau = \sum_{0}^{\infty} c_n \rho_1(n\tau_0) \quad (1$$

* Fitting the coefficients c_n so as to obtain the best least-mean-square approximati at given response function r has been discussed by Gabor.[9] Interesting corollar have recently appeared. K. L. Jordan (Quarterly Progress Report, M.I.T. Resear Laboratory of Electronics, April, 1959, p. 84) minimizes the truncation error; D. Tufts (*ibid.*, p. 87) minimizes the average error over a finite time interval, both appropriate choice of the interpolation functions. This can be done if the au correlation function of the process to be represented is known.

† VOLTERRA, V.: 'Theory of Functionals' (Blackie, 1930).

Here we have introduced a new response function ρ_1, which is the convolution integral of the filter response and of the first-order response h_1 of the general operator. By the theorem of resultants of Fourier integrals, the spectrum of this is the product of the frequency response $R(f)$ of the filter and of the Fourier transform $H_1(f)$ of $h_1(t)$.

The second term of eqn. (16) is similarly transformed into

$$\sum_0^{\infty}\sum c_{n1}c_{n2}\iint_{-\infty}^{\infty} r(t-\tau_1+n_1\tau_0)r(t-\tau_2+n_2\tau_0)h(t-\tau_1, t-\tau_2)d\tau_1 d\tau_2$$
$$=\sum_0^{\infty}\sum c_{n1}c_{n2}\rho_2(n_1\tau_0, n_2\tau_0) \quad . \quad . \quad . \quad . \quad (20)$$

This is a function of two time variables. It can be shown that its 2-dimensional Fourier transform is

$$R(f_1)R(f_2)H_2(f_1, f_2)$$

where H_2 is the 2-dimensional Fourier transform of h_2. The rule is easily generalized.

We thus arrive at an expansion for the output $g(t)$, modified by substituting for $f(t)$ its cardinal function, in the form of a series:

$$[g(t)] = \sum_0^{\infty} c_n\rho_1(n\tau_0) + \sum_0^{\infty}\sum c_{n1}c_{n2}\rho_2(n_1\tau_0, n_2\tau_0) + \ldots \quad (21)$$

This is a polynomial in the coefficients c_n, which are the expansion coefficients of the signal in terms of the response of the filter or, more simply, the samples of the pre-emphasized signal. The response functions ρ are required only at the Nyquist intervals, and hence in the text we have introduced for them the simpler notation $\rho_k(n_1\tau_0, n_2\tau_0, \ldots) = r_{n1\ldots nk}$. These coefficients are a fixed property of the operator; the time-dependence in eqn. (21) is contained in the c_n. To make this clear we have introduced in the text the notation f_n for these, which means 'a sample of the input taken at a time $t - n\tau_0$'.

(9.2) Optimization of a Non-Linear Filter

The most general problem which one can give to a filter is to produce a target function $\theta(t)$ by operating on the past of a time function $f(t)$. We can assume without loss of generality that the mean values of $f(t)$ and of $\theta(t)$ are both zero.

We form the error at time t:

$$O\{f(t)\} - \theta(t) = \sum f_n r_n + \sum\sum f_{n1}f_{n2}r_{n1n2} + \sum\sum\sum f_{n1}f_{n2}f_{n3}r_{n1n2n3} + \ldots - \theta_0 \quad (22)$$

The summations are to be carried out over all terms for which provision is made in the filter. As summation is to be carried out over all repeated suffixes the summation signs could be omitted, but we retain them for the sake of clarity.

The mean-square error, formed over a long time is (using the bar notation and the $\langle\,\rangle$ notation, as convenient)

$$\overline{[O\{f(t)\}-\theta(t)]^2} = \langle\sum\sum f_{n1}f_{n2}r_{n1}r_{n2} + \sum\sum\sum f_{n1}f_{n2}f_{n3}r_{n1}r_{n1n2} + \sum\sum\sum\sum f_{n1}f_{n2}f_{n3}f_{n4}r_{n1n2}r_{n3n4} + \ldots - 2\theta(\sum f_n r_n + \ldots) + \theta^2\rangle \quad (23)$$

We now introduce the following notation:

$$\overline{f_{n1}f_{n2}\ldots f_{nk}} = \phi_{k-1}(n_1 - n_k, n_2 - n_k, \ldots n_{k-1} - n_k) \quad (24)$$

This is the autocorrelation function of $f(t)$ of order $k-1$, defined for integral values of the arguments only. k is the number of variables in ϕ_k. With this definition the ordinary autocorrelation function is of the first order. We also define the cross-correlation function of $f(t)$ and $\theta(t)$ of order k:

$$\overline{f_{n1}f_{n2}\ldots f_{nk}\theta_0} = \psi_k(n_1, n_2 \ldots n_k) \quad . \quad . \quad (25)$$

The cross-correlation function as usually defined is again of the first order. With these notations the mean-square error becomes

$$[O\{f(t)\} - \theta(t)]^2 = \sum\sum r_{n1}r_{n2}\phi_1(n_1 - n_2) + \sum\sum\sum r_{n1n2}r_{n3}\phi_2(n_1 - n_3, n_2 - n_3)$$
$$+ \sum\sum\sum\sum r_{n1n2}r_{n3n4}\phi_3(n_1 - n_4, n_2 - n_4, n_3 - n_4) + \ldots$$
$$- 2\theta[\sum r_n\psi_1(n) + \sum\sum r_{n1n2}\psi_2(n_1, n_2) + \ldots] + \theta^2 \quad . \quad (26)$$

This is a positive definite quadratic form of the variables $r_{n1\ldots nk}$. The minimum condition gives the equations by differentiation with respect to the first-order responses r_n:

$$\sum\phi_1(n_1 - n)r_{n1} + \sum\sum\phi_2(n_1 - n, n_2 - n)r_{n1n2} + \ldots = \psi_1(n) \quad . \quad . \quad . \quad . \quad (27)$$

to the second-order responses r_{nm}:

$$\sum\phi_2(n_1 - n, n_1 - m)r_{n1} + \sum\sum\phi_3(n_1 - n, n_2 - n, n - m)r_{n1n2} + \ldots = \psi_2(n, m) \quad (28)$$

and so on.

As the mean-square error is a positive definite function of the responses, these equations always have a unique solution. The optimum responses will be zero only if $f(t)$ and the target function $\theta(t)$ are absolutely uncorrelated, in which case all the right-hand sides of the equations vanish. In all other cases some improvement can be effected. If the target function is a single-valued functional of its argument, and if the filter contains enough terms to reproduce this, the mean-square error will be zero, i.e. the filter will organize itself as a perfect simulator.

Eqns. (27) and (28) show clearly the information material used in the optimization process. The coefficients at the left-hand side are the autocorrelation functions of the process $f(t)$. The first-order autocorrelation function has been the object of much study, but relatively little work has been done on those of a higher order. In brief, they will all indicate regular repetitive occurrences in the process $f(t)$. The first-order autocorrelation will have peaks (positive or negative) if there are diads of peaks with set distances and correlated signs frequently occurring in $f(t)$. The second-order autocorrelation, which is a function of two variables, will indicate the presence of regular triads of peaks, and so on (cf. Gabor[9]).

The right-hand sides of eqns. (27) and (28) contain the cross-correlation functions between $f(t)$ and the target function $\theta(t)$. The first-order ψ has peaks if a peak in $f(t)$ at time $t - n/2F$ is regularly followed by a peak in θ at time t. The second-order ψ will have a peak if a diad in $f(t)$ is regularly followed by a peak in θ, and so on. This makes it easy to understand how the filter can act as a recognizer or recoder. Assume, for instance, that there is a certain pentade, such as 1, −1, −1, 1, 1, regularly occurring in $f(t)$ which the target function recodes into, say, −1, +1. To do this it assumes the value −1 as soon as the whole pentade has slipped into the past, and the value +1 when it has receded one step further. In this case the fifth-order ψ must have two non-zero values; all the rest can be zero. This simple machine would leave diads, triads and tetrads unnoticed, though it could be cheated by heptades.

[The discussion on the above paper will be found overleaf]

DISCUSSION BEFORE THE MEASUREMENT AND CONTROL SECTION, 6TH DECEMBER, 1960, AND BEFORE THE NORTH-WESTERN MEASUREMENT AND CONTROL GROUP AT MANCHESTER, 17TH JANUARY, 1961

Prof. A. Tustin: This machine, when completed, can give that polynomial in the values of the ordinates of an input function which most nearly agrees with the ordinates of an output function: but further developments, both of the theoretical background and of practical facilities for flexible adaptation, will be required before it can be applied to any wide range of real problems.

Many, if not most, of the problems in the general class of filtering or prediction involve more than one known input or disturbance. It appears, although this is not mentioned in the paper, that the machine could, in principle, cope with multiple inputs, but the number of channels required would become large and this is one of several reasons for seeking to economize in their number.

One simple possibility of doing this may lie in using the machine rather differently from the manner described. To illustrate the point, consider an output, θ_o, related to an input, θ_i, by the simple linear equation $\theta_o = \theta_i/(1 + pT)$. The weighting function in this case is exponential, and ordinates of the input extending back for several times the time-constant T have significant weight.

If, however, one writes the same equation as $\theta_o + d\theta_o/dt = \theta_i$ and approximates to $d\theta_o/dt$ by the first difference of the ordinates of θ_o, or by a better approximation involving only three or four ordinates, one sees that the equation is equivalent to a linear relation between one ordinate of θ_i and only two, or a few, ordinates of θ_o. Thus, if the machine were arranged to relate an ordinate of θ_o to the one simultaneous ordinate of θ_i, together with a few preceding ordinates of θ_o, a very short series would suffice, giving great economy in the channels required. It appears that in almost all problems it would be better to apply the machine to find a relationship between ordinates of both output and input rather than to include one ordinate only of the output as described in the paper. This would be particularly important when the data available are limited to records of short duration.

Economy in the number of coefficients of higher-order terms is discussed and is illustrated in Fig. 6. But where it is known that the non-linearity present is of some particular form, e.g. some simple power law, it may be possible to say in advance which coefficients are relevant and to devise a more rational scheme for curtailment and economy.

The use of a polynomial in ordinate values involving higher-order terms as a means for expressing the dependence of one quantity on another is not generally familiar to engineers, nor is it easy from this form of relationship to deduce anything about the physical structure of the system. If the authors could clarify the relationship of this form to the more familiar ones it would be very valuable.

Can the machine be used to discover the weighting function of a feedback, so as to minimize the mean square error of the system to which the feedback is applied?

Can the authors suggest any method to avoid the error associated with the commencement of the records analysed when these are very short?

Could the machine be developed so that, instead of minimizing an error, it showed how some weighting function could be adjusted to maximize an evaluating function regarded as output? This is a fundamental problem in plant control.

Mr. R. H. Tizard: Although named a filter, predictor and simulator, the machine described is fundamentally a parameter-estimator, for in the learning process it is estimating the parameters of the functional relationship between output and input. It could, in principle, be used as a filter, predictor or simulator, but in practice it is likely to be much too costly for such a use: it is more probable that it would be used to determine the settings for a filter, predictor or simulator having fixed or manually adjustable parameters.

Considering it, therefore, as a parameter-estimator, there are two important questions which the paper leaves partly or wholly unanswered. The first is the theoretical justification for the criterion of optimization used, and the second is the practical question whether the method adopted will achieve results within a practicable time in a machine of practicable size.

The authors state that from a practical point of view there is only one choice, the criterion of least mean squares, and they give no further justification. If the purpose of the machine were simply to achieve for a given input the nearest approach to a given output, one would be at liberty to choose any criterion of success, and the least-squares criterion would be as good as any other. But a machine of this limited purpose would have very little use, and one must assume that the real purpose consists in considering the given input and output as samples from a process which has some kind of statistical consistency, and estimating parameters which will give best results, on average, over other samples from the same process.

In these circumstances there is a proper theoretical justification for the least-squares criterion, for it corresponds, in linear processes at least, to parameter estimation by the method of maximum likelihood. This method, which has been extensively studied by statisticians, consists in finding that set of parameter values which maximizes the probability of having obtained those samples which were actually obtained. The factor whose mean square is to be minimized is that of 'uncorrelated residuals' in a discrete process, which corresponds to white noise in the continuous case, and, strictly speaking at least, it should be Gaussian.

The factor whose mean square is minimized by this machine, however, will not always be the right one. This may be illustrated by supposing that the two samples are from the input and output of a linear process subject to noise. If white noise enters the system only additively to the output, the criterion used will be correct. This case, however, is unlikely to occur except when the 'noise' is due simply to errors of measurement. In physical systems the noise enters at a number of points, and in this case minimizing the mean square of the difference between input and output will give the wrong results.

The correct least-squares may, however, then be obtained by operating on the output as well as the input. This reinforces the point made by Prof. Tustin. It is now apparent, however, that assumptions must be made, not only about the statistical nature of the noise, but also about the form of the process; for otherwise it is not possible to apportion the parameters between input and output. This illustrates the general point that a 'black-box' approach, by which the machine blindly adjusts itself to any input and target functions, will not do. It is absolutely necessary to have some *a priori* knowledge of the form of the process whose parameters are to be estimated.

Turning to the second point, the time taken by such a machine to reach the minimum, I cannot understand the statement in Section 2 that '. . . if there are M adjustable parameters the number of runs required for approaching the optimum to within a small fraction will be of the order $\frac{1}{2}M^2$'. This would mean only two runs for two parameters. In the second example of

the paper, with six parameters, 18 runs should have sufficed, yet the authors state that the minimum had not been reached after 230 runs. The characteristic of Fig. 13, of a rapid initial approach to the minimum, followed by a very slow convergence, seems to be common to optimum-seeking methods of this type. Such methods are, after all, only a form of numerical analysis, and possibly there are others which would converge more rapidly.

The authors mention that Dr. Wilkes has pointed out that 100 simultaneous linear equations can be solved in 7 minutes on the Edsac digital computer. They reject using such a computer, on the grounds that their own machine would have to be made more complex to obtain the necessary auto- and cross-correlation functions. But why not do this also by digital computer?

Mr. G. Pask: Will the authors comment on the relation between this learning machine and the brain, or one of the more brain-like artefacts?

There are occasions when a brain does much the same job as the predictive filter but, by comparison, the workings of the brain are untidy, information is lost and generalizations are made. In the animal (though, as Minsky and Selfridge[A] point out, not always in the artefact) these are cogent generalizations, and if the process is judged with reference to criteria which lay emphasis upon a particular image of the world (accenting, for example, those features pertinent to survival) animal learning is very efficient.

For all the obvious differences between this filter and the brain there are certain points of functional similarity. Consider, first, the kind of artefact called an 'abstractive filter'[B,C] which, in its simplest form, extracts certain 'attributes' of the sensory input. Structures of this sort have been identified in the frog's visual system[D] and seem to exist even in the primate brain. The categories specified in terms of these attributes have been extracted more efficiently than the categories of a logical network for two reasons—the attributes are relevant rather than exhaustive, and the abstractor is active (feedback pathways to the frog's retina) or adaptive (discrimination learning of phonemic categories in terms of a man's speech-sound attributes). The present filter 'abstracts' in so far as it weighs and sometimes neglects samples of past evidence. The rule for excluding or assigning coefficients and hence defining the relevant relations between the evidence is an inbuilt constraint. In common with the brain the efficiency of this device depends upon a nice balance of constraint and adaptation.

Next, consider one of the 'imitative' artefacts, such as MacKay's.[E] This machine, which represents the environment in terms of its own response set, is in many ways the opposite of the abstractive systems. Now, if we take a network of artificial neurones, an active one like that of Beurle,[F] with use-dependent reinforceable connections, there is some evidence that the system, in contact with a statistically stationary environment, will approach a structure which behaves either as an 'imitative' machine or like an abstractor. Using the word in Ashby's[G] sense these are stable 'partitionings' of the system. Which of the alternatives is preferred depends, amongst other things, upon the number of independent data channels connecting the artefact to its environment. At this level of discourse are there conditions which would lead a malleable network from initial homogeneity to a structure which 'worked like' the authors' filter? The papers by Greene[H] and Goodall offer one, as yet incomplete, view.

References

(A) Minsky, M., and Selfridge, O.: 'Learning in Random Nets', Proceedings of the 4th London Symposium on Information Theory (to be published).

(B) McCulloch, W. S., and Pitts, W.: 'How We Know Universals: the Perception of Auditory and Visual Forms', *Bulletin of Mathematical Biophysics*, 1947, **9**, p. 127.

(C) Von Foerster *et al.*: 'Some Principles of Preorganization for Self-Organizing Systems', Technical Report No. 2, Contract NONR 1834/21, University of Illinois.

(D) Lettvin, J. Y., *et al.*: 'What the Frog's Eye Tells the Frog's Brain', *Proceedings of the Institution of Radio Engineers*, 1959, **47**, p. 1940.

(E) MacKay, D. M.: 'The Epistemological Problem for Automata', Automata Studies, Study No. 34 (Princetown University Press, 1956).

(F) Beurle, R. L.: 'Properties of a Mass of Cells Capable of Regenerating Impulses', Philosophical Transactions of the Royal Society, B, 1956, **240**, No. 669, p. 55.

(G) Ashby, R.: 'An Introduction to Cybernetics' (Chapman and Hall, 1956).

(H) Greene, P.: 'A Suggested Model for Information Representation in a Computer Able to Perceive, Learn and Reason', *General Systems Yearbook*, 1961, **5**.

Dr. W. K. Taylor: I do not think the ear-plus-brain filter can perform the separation of music or conversation from a noisy background in terms of waveforms as the authors imply. If this were so we should not need high fidelity. What seems to happen is that in spite of the noise we are still able to classify and recognize the familiar patterns of speech and music, and this is a much simpler operation involving selection of the best match contained in our memory.

Regarding the speed of operation of the optimization process, the method of going through the parameters one at a time from 0 to 95 does not appear to be the best, and with reasonably simple additional equipment they could be taken in order of magnitude. The effectiveness of each parameter could be stored as a voltage on a capacitor during the first run. Each capacitor would be connected through a cathode-follower to a relay and through the relay to a common slowly-decreasing potential. The relay connected to the highest capacitor voltage would close first, discharging its own capacitor and returning the common potential to zero in readiness for the next run-down and the selection of the second largest capacitor voltage.

A number of workers in the field of automatic control have proposed simultaneous optimization procedures using independent noise generators to perturb the parameters and obtain cross-correlation signals to change all parameters at the same time. Do the authors think that this would lead to quicker optimization than the sequential method? Presumably, the parameter which has a big effect would start to act immediately, whereas it may be at the end of the sequential process.

What machine performance could be expected for a very simple input consisting of band-limited Gaussian random noise of zero mean? If the machine attempts to predict ahead as far as the next Nyquist sampling interval the most probable value is zero and all parameters should become zero, irrespective of their initial values. If the prediction time is reduced to zero, the first parameter, R_0, should become unity. What value does R_0 take for prediction times between zero and one Nyquist sampling interval? We know it should change from unity to zero, but the theory does not appear to indicate the intermediate values.

What happens if the input is like the random square wave $g(t)$ at the top of Fig. 1? If changes from $+1$ to -1 can only occur once per second the best prediction of the next state, using the mean-square-error criterion, is zero. If, however, you wish to maximize the probability of zero error it is better to

predict $+1$ or -1 continuously and be right half the time instead of never right. The theory on which the paper is based would give poor results, since the waveform contains very high frequencies and the sampling intervals tend to zero. In this case would it not be better to estimate the statistics of the sequence generator by detecting zero crossings and computing probabilities of temporal patterns?

Mr. J. L. Russell (*at Manchester*): Dr. Wilby gave an application for the use of the non-linear filter in a closed-loop system where it was necessary to provide exact correspondence between the input and output. There are methods of achieving this by synthesis of the input and the feed forward of this signal into the closed-loop control amplifier so as to provide exact correspondence.

A particular example is in position-control servo mechanisms where it is desired to run the servo with zero acceleration error. Such servos are first designed to run as zero-velocity-error systems, and a signal proportional to the acceleration of the input is synthesized from error and output position.

As $\theta_i = \theta_o + e$, $p^2\theta_i$ can be obtained by differentiating $\theta_o + e$ twice. This quantity, with a constant k, is fed forward and adjusted so as to zeroize the acceleration error. Additional terms, such as p^3, can be obtained by further differentiation of the synthesized signal. In practice, acceleration errors of $1'$/deg/sec^2 can be reduced to a tenth of this value. This attenuation can be achieved over the bandwidth of the system. Jitter which is present in the transmission system can be filtered by suitably-designed circuits interposed between the synthesizing amplifier and the servo amplifier.

Mr. J. D. Roberts (*at Manchester*): As three possible aspects of further simplification of the equipment: (*a*) Could a cruder and more inaccurate 'qualitative' multiplier be substituted, provided that it was consistent from run to run, and noting, for instance, that non-linearities in the first-order elements would be compensated by higher-order elements in the learning process? (*b*) Could a simplification be made in delays on a similar argument, using electronic delays in which some time 'blurring' was allowed? (*c*) Have the authors considered alternative methods of storing the coefficients, e.g. electrochemical storage?

In the field of applications, have the author's any estimate of the size of such a machine required to filter speech from noise effectively?

Mr. H. B. Daniels (*at Manchester*): The relationship between physical and human learning is brought into focus by the authors. The terminology used for the electronic simulator is the same as that used in the study of human psychology, and here lies a danger of assuming that the conclusions drawn from the electronic predictor apply to human beings. Can an electronic simulator have a 'memory' or only a 'store', 'sensory perception' or only an 'input programme', 'ear-plus-brain' filter or a 'sinusoidal signal' filter? Clearly, human aspects may be simulated only in their logical aspects and the effects of human incentive, imagination and emotion can be shown only as an 'input' to the logical electronic process of the simulator: beyond this, care is essential to avoid confusion as to what is physical and what human. Fig. 13 is a graphic illustration which the authors might indicate as being comparable to logarithmic growth in human learning.

Professor D. Gabor and **Drs. W. P. L. Wilby** and **R. Woodcock** (*in reply*): In reply to Professor Tustin, the machine is so designed that its input can be divided into three or six channels. The suggestion that simpler relations might be obtained if the machine were used for producing several output ordinates instead of one is certainly interesting, and the machine is well adapted to it. For lack of space, we have not discussed its use as a recognizer or coder. In such an application, for instance, it will definitely be better used if, after recognizing a code-sign, which is a group of amplitudes, it supplies at once the corresponding group of amplitudes instead of giving ordinates one by one.

Though polynomial representations of non-linear processes with memory may not be familiar to engineers, they are well known to mathematicians, as explained in connection with eqn. (16).

It is quite straightforward to use the machine for discovering the weighting function of a feedback. All one has to do is to insert the filter as a feedback link into the closed-loop system. This is also in reply to Mr. Russell. The analysis of short statistical series is outside our line and would require much special study. This is also partly in reply to the comments of Mr. Tizard. It is expecting a little too much of our machine that it should be applicable to the completely different problem of maximizing an output function, but it is very likely that much of our detail work would prove useful if such a machine were to be constructed.

Mr. Tizard's comments open up a wide field which we do not wish to enter. As regards $\frac{1}{2}M^2$, it is a rule of thumb which states that a 'satisfactory' approximation is reached after this number of runs. In Fig. 13, for instance, after 18 runs the error was within 10% of the minimum. We have given much consideration to the question whether the whole machine or part of it should be digital, but comparisons of expenditure always made us prefer the analogue realization.

Mr. Pask's comments also open up a wide field of speculation. Cautiously, one can only say that our machine is as different from a brain as possible. It is quantitative instead of classificatory, with a precise quantitative memory and no sense whatever for generalization. If one wished to imitate the action of human or animal intelligence, it would be better to make a fresh start rather than to try to construct a bridge from our machine to a nervous system.

This also covers the first remark of Dr. Taylor. He may be wrong in thinking that the ear-plus-brain filter cannot perform the separation of music and noise; anybody can observe that, after playing a gramophone record many times, the noise becomes less and less. But he may well be right in thinking that the filter action is quite a different process in the two cases. While the machine constructs the purified music piecemeal from the noisy original, the trained human memory in the end just uses the senses for the perfunctory recognition and pulls the rest out of the memory where it is preserved in a purified form. We have given much consideration to improved methods of descent such as Dr. Taylor suggests and, in fact, we are applying these, but at the present stage there appears no justification for mechanizing them. In the somewhat extreme example which he mentions, it might be indeed better to detect zero crossings. In general, it pays to do some thinking oneself and to extract as many of the regularities as possible in advance, before setting the machine to the task.

In reply to Mr. Roberts, the application of this machine to the filtering of speech from noise is almost hopeless, because the way from waveforms to phonemes is very long. One could, however, think of using it as a filter once we have learnt how to abstract the six essential speech parameters of W. Lawrence; it would then smooth and predict these parameters instead of the raw amplitudes.

In reply to Mr. Daniels, we have used the evocative vocabulary of machine builders without much compunction. Anybody who has qualms can retranslate 'memory' into 'store', 'sensory perception' into 'input programme'.

A SELF-OPTIMIZING NON-LINEAR FILTER

By J. K. Lubbock, Ph.D.

(*Communication received 25th January*, 1961.)

Self-optimizing procedures may be employed advantageously for the solution of three distinct types of problem:

(*a*) The solution of a set of equations which minimize some 'cost' function. This is done in the design of some chemical plants where the cost function may depend upon the quantity and quality of output product, capital investment and running costs.

(*b*) The adjustment of an existing physical system to improve its response, where certain parameters or signals can be altered, but in the main the plant is of fixed design. Problems of this kind arise in finding the best operating conditions of an existing chemical plant.

(*c*) The self-adjustment of a filter so that it can perform some specified operation such as prediction or simulation, where the design of the filter is left entirely to the discretion of the designer.

In the first two cases, hill-climbing manoeuvres will inevitably be needed, but the third type of problem, with which the paper by Professor Gabor and Drs. Wilby and Woodcock (see p. 422) is concerned, is capable of a simpler solution since it is possible to design the filter so that hill-climbing manoeuvres are not needed.

In Section 2 the authors dismiss the use of orthogonal polynomials without giving sufficient justification for their action. The use of orthogonal functions, both linear and non-linear, for this type of self-adjusting filter has great advantages; this will now be demonstrated.

Orthogonalization of the terms of eqn. (4) of the paper from left to right, exhausting all those terms under the first summation before proceeding to the second, etc., will yield a set of orthogonal polynomials, the first $N + 1$ of which involve only linear terms:

$$\left.\begin{aligned} \theta_0 &= f_0 \\ \theta_1 &= f_1 - a_{10}\theta_0 \\ \theta_2 &= f_2 - a_{21}\theta_1 - a_{20}\theta_0 \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \theta_{N+1} &= f_1 f_2 - a_{N+1,N}\theta_N - \cdot \cdot \cdot \\ &\text{etc.} \end{aligned}\right\} \quad . \quad . \quad (1)$$

where

$$\begin{aligned} \overline{\theta_n\theta_m} &= 0 \quad (n \neq m) \\ &= A \quad (n = m) \quad . \quad . \quad . \quad . \quad (2) \end{aligned}$$

A is a constant and the bar indicates a time average. From this it is easily proved that

$$a_{nm} = \frac{\overline{f_n\theta_m}}{\overline{\theta_m^2}} \quad (n \leqslant N) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

$$a_{nm} = \frac{\overline{f_n f_v \theta_m}}{\overline{\theta_m^2}} \quad [N < n \leqslant \tfrac{1}{2}(N+1)(N+2)] \quad . \quad (4)$$

etc.

Thus, all a_{nm} can be measured experimentally, taken in order starting with a_{10}, by synthesizing each orthogonal polynomial in turn (after having determined its coefficients) and using its output for the measurements of the coefficients of the next orthogonal polynomial. a_{nm} can also be set by a self-adjusting procedure to give a null in the measured value of $\overline{\theta_n\theta_m}$; again the coefficients must be taken in order starting with the lowest, but all the coefficients belonging to any one orthogonal polynomial can be set simultaneously without interaction, which means there will be only $M - 1$ distinct operations in the setting-up procedure of M polynomials.

The orthogonal functions having been determined, the coefficients of the linear summation, $\sum_n b_n\theta_n$, for the optimum filter can be found very easily without employing hill-climbing manoeuvres, since each b_n is given by the independent algebraic equations which are obtained by minimizing the mean-square error:

$$\overline{e^2} = \overline{\left\{g(t) - \sum_n b_n\theta_n[f(t)]\right\}^2} = \overline{g^2} - 2\sum_{n=0} b_n\overline{g\theta_n} + \sum_{n=0} b_n^2\overline{\theta_n^2} \quad (5)$$

where other terms are zero because of the orthogonal property of eqn. (2); $g(t)$ is the target signal. For minimum mean-square error,

$$\frac{\partial\overline{e^2}}{\partial b_n} = 0 = -2\overline{g\theta_n} + 2b_n\overline{\theta_n^2}$$

Therefore

$$b_n = \frac{\overline{g\theta_n}}{\overline{\theta_n^2}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The coefficients b_n can be measured, or set by a self-adjusting procedure to give a null in the measured value of

$$\overline{\{g(t) - b_n\theta_n[f(t)]\}\theta_n[f(t)]} \quad . \quad . \quad . \quad . \quad (7)$$

These adjustments may be made simultaneously because each function θ_n makes an independent contribution to the reduction of mean-square error. This adjustment constitutes one further operation.

Therefore the whole filter can be optimized in M distinct operations, and without noise an exact optimum is obtained, whereas the authors' hill-climbing procedure produces an asymptotic convergence to the summit so that the exact optimum is never reached.

If the sum total of the a_{nm} in any one orthogonal function is adjusted simultaneously and the sum total of the b_n is adjusted simultaneously, the equipment will need M multipliers and integrators for the self-adjusting circuits. Using only one multiplier and one integrator and setting all the coefficients in sequence necessitates a total number of operations given by

$$\begin{aligned} \sum_{r=0}^{M-1}(M - 1 - r) + M &= M^2 - \sum_{r=0}^{M-1} r \\ &= M^2 - \frac{M(M-1)}{2} \\ &= \tfrac{1}{2}M(M+1) \end{aligned}$$

Thus, the greatest number of operations required to reach an exact optimum is insignificantly greater than the number which is estimated by the authors to be necessary for their hill-climbing machine to reach a near optimum. In the first problem of the paper (Section 5.9), with two adjustable parameters, three adjustments would give an exact optimum for the orthogonal filter, whilst after three adjustments the hill-climbing machine still shows about a 7% mean-square error.

There is therefore little to lose by employing orthogonal polynominals since

(a) An exact instead of an approximate optimum is obtained and there will be no indecision on how long to let the machine run.
(b) The number of operations is never more than that required to adjust interacting parameters.
(c) By using more equipment much faster optimization can be accomplished.

On the debit side, it must be admitted that more voltage dividers must be used, but to offset this there is no need to employ mechanical strategies (see Section 4 of the paper).

It is appropriate to remark that the complexity of filters employing perception (as the authors call it) can be very great indeed. If the past of the input is characterized by N past sample values and the filter employs only first-order perception [using only those terms under the single and double summations of eqn. (4)] we shall need

$$\sum_{r=1}^{N+1} r = \tfrac{1}{2}(N+1)(N+2) \text{ multipliers.}$$

For filters employing second-order perception we shall need an additional number of multipliers equal to

$$\sum_{s=1}^{N+1} \sum_{r=1}^{s} r$$

Likewise for Kth order perception we need

$$\underbrace{\sum_{n=1}^{N+1} \cdots \sum_{s=1}^{p} \sum_{r=1}^{s} r}_{K} \text{ additional multipliers.}$$

The total number of multipliers for Kth order perception is

$$X = \sum_{k=1}^{K} \Big(\underbrace{\sum_{u=1}^{N+1} \cdots \sum_{s=1}^{p} \sum_{r=1}^{s} r}_{k} \Big)$$

This expression can be simplified by writing

$$X = \sum_{k=1}^{K} \left[\frac{(N+1)(N+2)\ldots(N+k+1)}{(k+1)!} \right]$$

which gives, for example, the following values for X:

K	N	X
2	2	16
2	18	1 520
3	18	8 835
7	18	2 220 055

It is seen that the number of multipliers becomes prohibitive except for very small values of K and N. It seems, therefore, that the suitability of simpler classes of filters should be investigated; in this way more practical optimum filters will be produced.

One simplification which can be made to the authors' filter without loss of generality is to use a different method of characterizing the past of the input so that N can be made smaller. The most concise method is to use the phase co-ordinates[2] of the input; another method is to use Laguerre functions[3] of the past or some other suitable set of functions which can be chosen to be orthogonal with respect to the input spectral density.[4,5] All of these methods employ linear networks for the generation of f_n, each of which is dependent upon the whole past in a smoothly weighted fashion; thus it is to be expected that fewer terms may be needed for the same filter performance than when sampled values are used. (It can be shown that this is true using the phase co-ordinates for Gaussian inputs.) The use of the correct orthogonal functions for characterizing the past has the added advantage that the cross-correlation of their outputs is zero, so that fewer multipliers are needed in the generation of the non-linear terms having first-order perception.

References

(1) Lubbock, J. K.: 'Optimization of a Class of Non-Linear Filters', *Proceedings I.E.E.*, Monograph No. 344 E, November, 1959 (**107** C, p. 60).
(2) Fuller, A. T.: Ph.D. thesis, University of Cambridge.
(3) Wiener, N.: 'Extrapolation, Interpolation and Smoothing of Stationary Time Series' (Wiley, 1949).
(4) Kitamor, T.: 'Applications of Orthogonal Functions to the Determination of Process Dynamic Characteristics and to the Contraction of Self-optimizing Control Systems', Proceedings I.F.A.C. Congress, Moscow, 1960 (Butterworths).
(5) Lubbock, J. K.: 'Mathematical Models of Plants', Proceedings Seminar entitled 'Le calcul analogique appliqué a l'étude des processus chimiques', Brussels, 1960.

[Professor D. Gabor and Drs. W. P. L. Wilby and R. Woodcock make the following comments:

We fully recognize, of course, that the hill-climbing manoeuvre in our case is really the solution of a system of simultaneous linear equations. For reasons of economy, we prefer a hill-climbing (or rather, descending) procedure because, in order to do this, we need not know at any time more than a certain section of the hill, and this information is supplied by much simpler apparatus than that required for the realization of Dr. Lubbock's proposals.]

21—526

The Institution of Electrical Engineers
Paper No. 3523 M
July 1961

A SELF-OPTIMIZING NON-LINEAR CONTROL SYSTEM

By J. L. DOUCE, Ph.D., M.Sc., and R. E. KING, Ph.D., M.Sc., Graduates.

(*The paper was first received* 2*nd September, and in revised form* 12*th December*, 1960.)

SUMMARY

A general technique of self-adaptation is presented which adjusts one parameter of a control system to obtain the best response as determined by some built-in optimizing criterion. The method has been applied to a position-control servo mechanism in which the damping factor is automatically adjusted to give minimum mean squared error. The response of the control system is considered for repetitive step functions and sinusoidal and random signal inputs.

The addition of the self-adjusting feature requires little knowledge of the controlled process dynamics or applied signal characteristics. The optimization mechanism compensates for variations in system parameters, non-linear effects and changes in the applied signal.

It is shown that the introduction of a self-optimizing element effects a substantial improvement in system performance. Experimental results using an analogue computer are in good agreement with theoretical predictions.

LIST OF PRINCIPAL SYMBOLS

x_i = Input signal.
x_o = Output signal.
$x_e = x_i - x_o$ = Error signal.
x_c = Control signal.
σ^2 = Mean squared power.
$\Phi_i(j\omega)$ = Input r.m.s. voltage spectrum.
$\Phi_i(0)$ = Input zero-frequency r.m.s. voltage per unit bandwidth.
$\omega_0 = 1/T$ = Control-system undamped resonant frequency.
$W(p, \sigma_i)$ = Closed-loop transfer function.
$E(p, \sigma_i)$ = Error-ratio transfer function.
$C(p, \sigma_i)$ = Control-ratio transfer function.
μ = Ratio of control system resonant frequency ω_0 to input spectrum half-power frequency.
K_t = Booton's equivalent gain for torque limiter.
K_v = Control-system damping factor.
ΔK_v = Damping-factor perturbation amplitude.
ψ_v = Damping-factor misadjustment.
e_M = Error measure of performance.
ω_s = Perturbation frequency.

An asterisk (*) denotes an optimum value.

(1) INTRODUCTION

The ultimate aim in the design of a servo mechanism subjected to a signal uncontaminated by noise is, in general, to minimize or maximize a given performance index or criterion, which is some function of error (the difference between input and output). An 'optimum' system can be defined as one which, by suitable adjustment of its parameters, satisfies these requirements.

For a linear system the optimum setting of any parameter is unaffected by variations in input-signal amplitude, and depends only on the frequency spectrum of the applied signal. In many cases, infinite loop gain provides minimum or zero error and the linear equation may no longer accurately represent the practical system.

When inherent limitations on the maximum signal amplitudes are considered, a more realistic optimum system may be designed. The variable parameters of the system must now be adjusted for each change in input amplitude or spectrum, and owing to the complex mathematical analysis involved, it is usual for only one parameter to be adjusted. If the characteristics of the input signal are known, the amplitude-dependent parameter variation may be incorporated as a non-linear element. It has been shown that a non-linear error detector incorporated in a saturating position-control system gives the optimum step response for all input amplitudes[1] and an enhanced performance for random signal[2] and sinusoidal inputs.[3]

The inclusion of some specified parameter variation of this type comes under the category of 'passive adaptation'. Adaptive control systems can, in general, be divided into two broad categories, passive and active, and are distinguished by their ability to self-compensate for two basic types of disturbance, namely:

(*a*) Changes in system input characteristics: these can be in the form of changes in power, spectrum, signal/noise ratio and load fluctuations.

(*b*) Changes in system parameters: examples of these are component variations and environmental changes.

In the class of passively adapted systems, in which 'optimum' performance is claimed for a particular system, it is immediately implied that the designer has an *a priori* knowledge of all possible input conditions and process dynamics. In many instances, such supposition of constant spectrum, power and amplitude probability of the input signal may not be valid, and hence for optimum performance this form of system compensation is impossible to evaluate.

In other instances the characteristics of the input signal to the control system might be known to be changeable and possess non-stationary properties, or be altogether unknown, even statistically. In such cases, therefore, passive adaptation cannot be considered to possess characteristics compatible with optimum performance. Only in such instances, or where the physical process is indeterminate and widely varying in its dynamic characteristics, is the use of a self-adaptive or actively adapted system justifiable. In this form of adaptation system, optimization is achieved by automatic parameter variations.

Thus, under circumstances of completely unpredictable environment in the form of either unknown input-signal spectrum and power or unknown dynamic characteristics of the controlled physical process, a self-adaptive system which is continually searching for the optimum operating conditions will give superior performance compared with any passive optimizing approach.

Examples of dynamic performance liable to variation, and for which, therefore, self-adaptive systems are eminently suited, are: aircraft and guided missiles, whose dynamic characteristics vary with speed and altitude; chemical processes; and large-scale control operations where the system dynamics are unknown and are liable to unforeseen changes.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Douce is and Dr. King was formerly in the Electrical Engineering Department, Queen's University of Belfast. Dr. King is now in the Autonomics Division of the National Physical Laboratory, Teddington.

In the system described, automatic parameter variation is achieved by a trial-and-error technique due to Draper and Li,[4] by which a parameter is varied, the effect noted and used to modify the parameter itself. The process is rendered continuous by introducing a cyclic parameter variation.

The method of adaptation proposed is applicable to a wide range of complex systems, and has been applied to a position-control servo mechanism in a manner similar to that proposed by Taylor.[11] After discussing the behaviour of the practical system, a non-linear second-order system with an adaptive loop is investigated on an electronic simulator, and theoretical derivation of the optimum conditions is compared with the observed operation.

(2) BASIC SYSTEM

Fig. 1 shows the basic system for automatic parameter variation. It is assumed that the measure of system performance

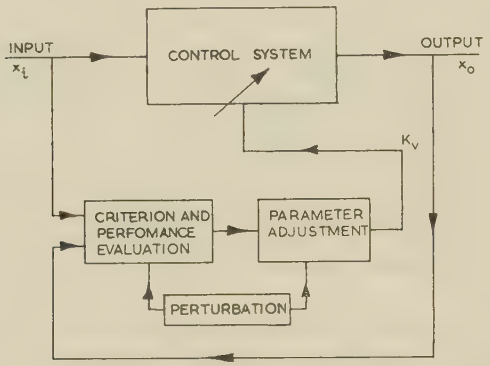


Fig. 1.—Adaptive system.

has been determined: this measure can be any instantaneous function of output or error. In response to a step change of some parameter, K_v, there will be, in general, a change in output and error. This change will consist of a transition from the initial operation to some new operating conditions. The transient response may be determined by considering the response of the new system to the values of error, error rate, etc., existing at the switching instant, giving the time required for the new steady-state condition to be attained. This response is the time response of the equivalent linear system subjected to initial values of error, etc., as given by the normal transient response of the system. Hence, to determine the effect on the steady-state performance of some parameter variation, it is necessary to measure over a time large compared with the response time of the system. When the parameter variation is introduced in a cyclic manner, the switching frequency must be much less than the natural frequency of the control system. This conflicts with the aim of rapid optimization, which requires high rates of searching, and a compromise must always be attained. Practical values for the switching frequency, ω_s, lie in the range $0 < \omega_s < \frac{1}{9}\omega_0$.

When K_v is varied in some cyclic manner, the error magnitude will contain a component varying in a similar manner, and it is necessary to extract information giving the better value of K_v, so that the mean of this parameter may be adjusted to some better value.

Fig. 2 illustrates the method adopted, in which the squared-error of the system is phase-sensitive rectified, using the parameter perturbation as a reference signal. This is achieved simply in practice, when the perturbation ΔK_v is a square wave, by multiplying the error by ± 1.

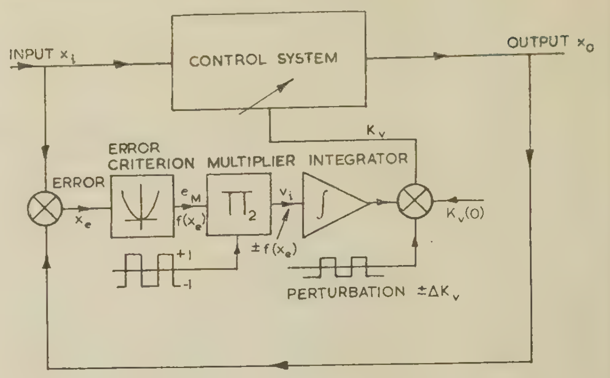


Fig. 2.—Particular system.

The magnitude of this signal is a measure of the effect of the parameter variation on performance, and the sign indicates the sign of ΔK_v corresponding to a reduction in error. Thus, if this signal is fed into an integrator, the integrated waveform is suitable for modifying the parameter K_v.

Hence the parameter K_v is composed of three elements: the initial value $K_v(0)$ which may be of any magnitude provided that the control system does not become unstable, a cyclic variation, of small magnitude, conveniently a step function waveform; and the output of the optimizing unit.

It will be noted that, in the Figure, the criterion chosen in the example—mean squared error—is taken to be a function of time. Here and throughout the paper, the mean-square value is taken as an ensemble average, rather than as a time average.

(2.1) Response of Adaptive Loop

In the general form of the adaptive system shown in Fig. 2 the adaptive loop includes two non-linearities, one of which is frequency-dependent, and a multiplier. However, since the frequency of operation of the adaptive loop is much less than the natural frequency of the control system, an approximate analysis may be achieved by assuming an instantaneous relationship between the parameter variation and the error measure e_M.

In general, considering the steady-state conditions,

$$e_M = \mathrm{f}(K_v)$$

with $$K_v = K_v(0) \pm \Delta K_v$$

i.e. $$e_M = \mathrm{f}[K_v(0) \pm \Delta K_v] = \mathrm{f}[K_v(0)] \pm \frac{\partial \mathrm{f}}{\partial K_v}\Delta K_v$$

After synchronous detection, this gives the signal at the input to the integrator

$$v_i = \frac{\partial \mathrm{f}}{\partial K_v}\Delta K_v$$

If the manner in which the mean squared error varies with changes in parameter is known, the response of the adaptive loop to step changes in parameter is given by

$$K_v(t) = K_v^* + \psi_v \delta_{-1}(t) - \frac{\Delta K_v}{T_i}\int_0^t \frac{\partial \mathrm{f}}{\partial K_v} dt$$

where ψ_v is the initial parameter misadjustment, T_i is the adaptive-loop gain and $\delta_{-1}(t)$ is the unit step function. K_v^* is the optimum value of parameter setting.

In the practical examples considered later, where the adaptive loop varies the amount of velocity feedback in a position-control servo mechanism, a typical relationship between error measure

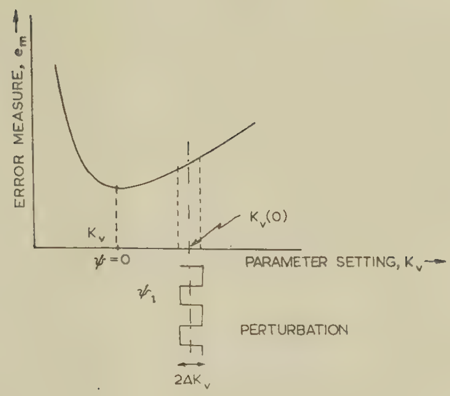


Fig. 3.—Static transfer characteristic.

(mean square) and parameter setting (damping) is shown in Fig. 3. It follows that two approximations to the non-linear characteristic $f(K_v)$ may be considered. For K_v greater than optimum, the relationship is assumed linear of slope S_1, so that

$$v_i = S_1 \Delta K_v$$

and for K_v less than optimum, a parabolic characteristic is assumed, i.e.

$$e_M = E + S_2(K_v^* - K_v)^2$$
$$= E + S_2\psi^2$$

so that $$v_i = 2S_2\psi\Delta K_v$$

For a positive misadjustment, therefore, the transient response of the adaptive loop is

$$K_v(t) = K_v^* + \psi_v\delta_{-1}(t) - \frac{\Delta K_v}{T_i}S_1 t$$

and for a negative misadjustment,

$$K_v(t) = K_v^* - \psi_v\delta_{-1}(t) - \frac{2\Delta K_v}{T_i}S_2\int_0^t \psi dt$$

yielding a first-order differential equation:

$$\frac{dK_v}{dt} + \frac{2\Delta K_v S_2}{T_i}K_v = -\psi_v\delta_0(t)$$

where $\delta_0(t)$ is the unit impulse function.

Thus

$$K_v(t) = K_v^* + \psi_v(1 - \varepsilon^{-2\Delta K_v S_2 t/T_i})$$

It follows that, if the initial parameter setting is too small, an exponential approach to the optimum is achieved, and, for initial settings greater than optimum, the parameter will approach the optimum linearly with time. This analysis has neglected the finite amplitude of the perturbation, which gives a slower ultimate rate of settling, and also the time response of the control system, which can introduce an oscillatory parameter variation if the frequency of perturbation is too high.

(3) A PRACTICAL SYSTEM

To investigate the practical application of the suggested method, an adaptive loop, consisting of a velodyne integrator multiplier and a relay demodulator, has been constructed and applied to a position-control servo mechanism in which the amount of velocity feedback may be varied.

The overall system is shown in Fig. 4. The general configuration of the optimizer is such that any performance criterion which can be expressed in the form

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \phi[x_e(t)]dt$$

can be applied. This is achieved, in practice, by feeding the error signal into an assessing unit or 'criterion function' which follows the desired law. As minimum mean squared error has been taken as the desired optimum operating condition, the criterion function is of the form

$$\phi[x_e(t)] = Ax_e^2(t)$$

and is generated electronically by a biased-diode square-law characteristic.

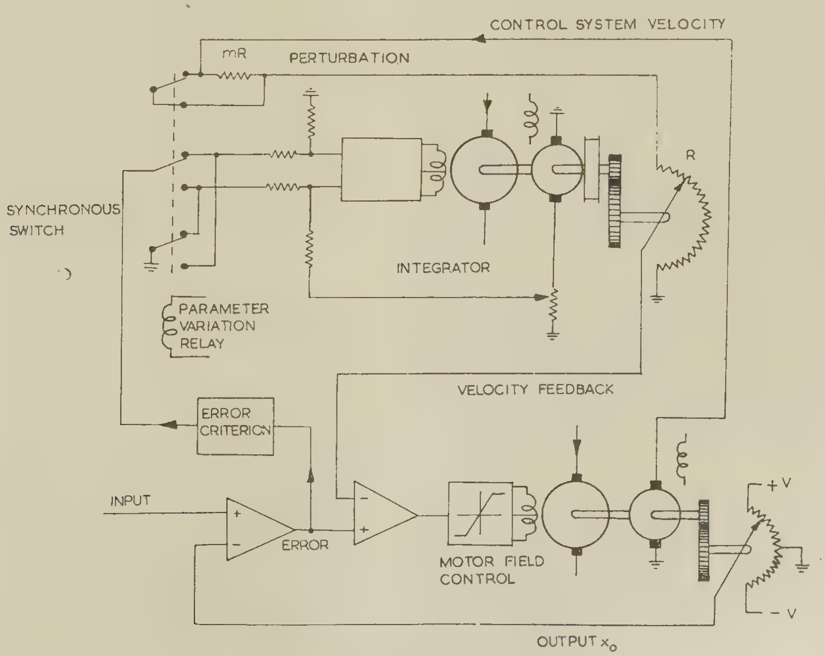


Fig. 4.—Practical system.

The input to the system is a random signal of approximately constant power per unit bandwidth over the range 0–0·25 c/s. For small signals, the undamped natural frequency of the system is 3 c/s.

The perturbation rate is 0·5 c/s, and since it is most important for the perturbation-signal square wave to be accurately unity mark/space ratio it is obtained from a scale-of-2 counter driving the relay.

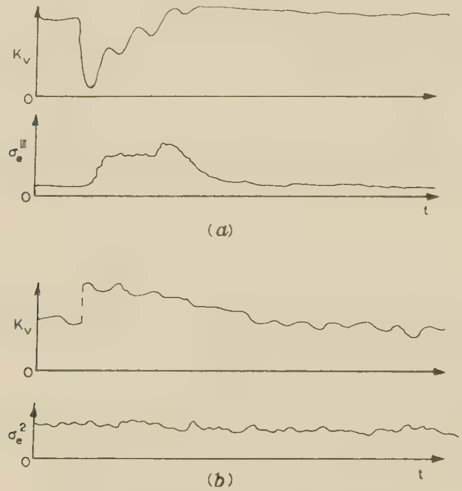


Fig. 5.—Adaptive loop response to step misadjustments in damping.
(a) $\psi < 0$.
(b) $\psi > 0$.

Fig. 5 shows the effect of sudden changes of the system damping factor on the mean squared error and the resultant changes in the value of the damping coefficient for both positive and negative misadjustments. This illustrates the rapid recovery when a large change in error is involved, and the slow optimization when the change in error is less significant.

(4) ANALOGUE-COMPUTER VERIFICATION

To obtain a quantitative assessment of the efficiency of the optimizing mechanism, a second-order saturating system with its adaptive loop has been simulated on an analogue computer. This simple system has been chosen for ease of analysis only and provides no fundamental limitation on the general technique.

The best system is defined as that possessing minimum mean-square error as the amount of damping is varied, and analytical expressions are developed for the best system when the input is a repetitive step function, a sinusoidal signal, and for Gaussian noise possessing a simple power spectrum. Experimental results are presented in each case.

Fig. 6 shows the block diagram of the system employed, with the velodyne integrator multiplier of the mechanical system replaced by an electronic drift-compensated integrator and an electronic multiplier. A high-speed relay was used for error sampling and parameter perturbation, and the technique is similar to that used in the practical system, but here

$$\frac{\Delta K_v}{K_v} = \frac{1}{2(1+m)}$$

where mR is the resistance determining the magnitude of the perturbation, as shown in Fig. 6. Hence $m = 4$ for a 10% perturbation.

The response of the overall adaptive system is considered for a variety of applied input signals, and it has been assumed that

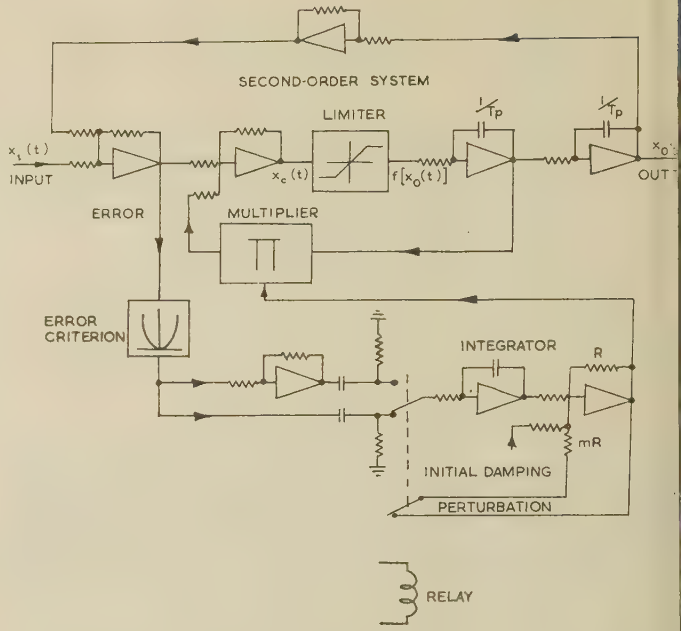


Fig. 6.—Analogue of system.

the dynamic behaviour of the system can be mathematicall expressed and is known.

First, the equivalent linear system is analysed, and the varia tion in integral-error squared as a function of damping factor determined for step inputs. The torque-limited adaptive syster is compared, and it is shown to possess characteristics of th Serme and $V|V|$ systems. The effect of a sinusoidal input o the adaptive system behaviour is then considered, and finall the more realistic case of a random signal input to the systen for both linear and non-linear operation is discussed.

(4.1) Transient Response

(4.1.1) The Linear System.

For a linear system characterized by the second-order transfe function

$$W_0(p) = \frac{1}{T^2p^2 + K_vTp + 1}$$

subjected to a step input

$$x_i(t) = N\delta_{-1}(t)$$

the integral-error squared is

$$I_e = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} |[1 - W_0(p)]x_i(p)|^2 dp$$

and, since $x_i(p) = N/p$,

$$I_e = \frac{N^2T}{2K_v}(1 + K_v^2)$$

Minimum integral-error squared occurs for that value of K_v fo which

$$\frac{\partial I_e}{\partial K_v} = 0$$

giving $K_v = 1{\cdot}0$ (i.e. half critical damping) and $I_e^* = N^2T$ The output for this optimum condition is then given by

$$x_o(t) = N\left[1 - \frac{2}{\sqrt{3}}\varepsilon^{-t/2T}\cos\left(\frac{\sqrt{3}}{2}\frac{t}{T} - \frac{\pi}{6}\right)\right]$$

.1.2) Torque-Limited System.

When large steps are applied to the system, so that the initial ror is much greater than that required for maximum torque, e motor will be working almost entirely under conditions of aximum torque giving constant acceleration or deceleration.

Previous work has shown that the fastest response is now oduced by applying maximum acceleration until the error is duced to one-half the initial value, followed by maximum eceleration, so that the system comes to rest with zero error. his mode of operation may be achieved by non-linear velocity edback (the $V|V|$ system[5]) or by a non-linear error detector he Serme system[1]).

It may be shown that this type of response corresponds to the est as determined by the minimum mean-squared-error criterion.

For a control system whose motion in the linear regime is overned by the second-order differential equation

$$T^2\ddot{x}_o + K_vT\dot{x}_o + x_o = 0$$

e slope of the trajectory on the phase plane is given by

$$\frac{d\dot{x}_o}{dx_o} = -\frac{x_o + K_vT\dot{x}_o}{T^2\dot{x}_o}$$

nd is zero when

$$x_o + K_vT\dot{x}_o = 0$$

hich is the equation for the change-over boundary.

When a large input step is applied to the torque-limited system, e maximum acceleration of the motor is

$$\ddot{x}_{om} = \frac{h}{T^2}$$

here h is the maximum motor torque. On integrating, the elocity is

$$\dot{x}_o = \frac{ht}{T^2}$$

ssuming that the system is initially at rest. The output signal then

$$x_o = \frac{ht^2}{2T^2} = \frac{T^2\dot{x}_o^2}{2h}$$

which is the characteristic parabolic form of the trajectory on he phase plane.

For correct change-over, i.e. zero overshoot, torque reversal nust take place when $x_o = \frac{1}{2}N$, at which point the equations of notion on the phase plane and change-over boundary must both old. To achieve the desired response, therefore,

$$K_v = \frac{1}{2}\sqrt{\frac{N}{h}}$$

Under such conditions the minimum integral-error squared is

$$I_e = \int_0^{T\sqrt{(N/h)}} \left(N - \frac{ht^2}{2T^2}\right)^2 dt + \int_0^{T\sqrt{(N/h)}} \left(\frac{1}{2}\frac{ht^2}{T^2}\right)^2 dt = 0{\cdot}766N^{5/2}Th^{-1/2}$$

On applying a repetitive step input of large magnitude to the daptive system, the damping factor, K_v, will so adjust itself hat the trajectory on the phase plane will meet the relevant hange-over isocline when the output signal becomes exactly alf the input magnitude. The choice of this change-over oundary is thus identical to the $V|V|$ and Serme systems.

The variation of damping factor with input step magnitude is hown in Fig. 7. For small inputs the damping factor will tend

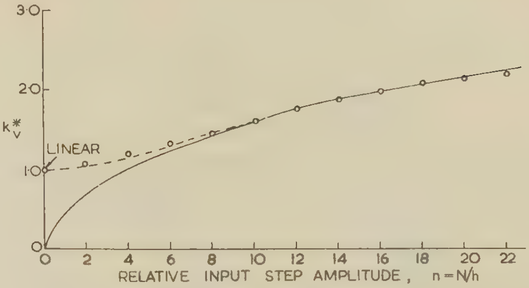


Fig. 7.—Optimum damping for torque-limited system for step inputs.

— Theoretical, neglecting motion in the linear regime.
○ Experimental.

to unity and not to zero, as given by the theory neglecting the motion in the linear regime.

(4.2) Sinusoidal Response

For a sinusoidal excitation signal

$$x_i(j\omega) = \theta_i\varepsilon^{j\omega t}$$

The error signal of the linear system will be

$$x_e(j\omega) = x_i(j\omega)[1 - W_0(j\omega)]$$

The peak-error-squared ratio is thus

$$\left|\frac{\hat{x}_e(j\omega)}{\hat{x}_i(j\omega)}\right|^2 = E_0^2(\alpha) = \frac{\alpha^2 + K_v^2}{(1-\alpha^2)^2 + \alpha^2K_v^2}\alpha^2$$

where $\alpha = \omega T = \omega/\omega_0$ is the normalized input frequency. This relationship can be rewritten in the form

$$E_0^2(\alpha) = 1 - \frac{(1 - 2\alpha^2)}{(1-\alpha^2)^2 + \alpha^2K_v^2} = 1 - J(\alpha)$$

Now, for $\alpha < 1/\sqrt{2}$, $J(\alpha) > 0$, and therefore it is desirable for minimum mean squared error to have $J(\alpha)$ as large as possible by having its denominator as small as possible, implying $K_v = 0$. Conversely, for $\alpha > 1/\sqrt{2}$, $J(\alpha) < 0$, and therefore it is now desirable to have $J(\alpha)$ as small as possible, i.e. $K_v = \infty$.

This solution leads to the concept of an idealized low-pass filter, whose characteristics are portrayed in Fig. 8. The phase of the transfer function is zero in the pass-band and indeterminate outside this. It is important to note that this characteristic does not apply when the input signal contains, simultaneously, components of frequencies greater and less than $\omega_0/\sqrt{2}$.

On subjecting the particular adaptive system to a sinusoidal input signal of frequency less than the critical frequency, zero damping is demanded, but as soon as this condition is met the control system bursts into oscillations of resonant frequency ω_0. As soon as these oscillations begin to build up, the error signal rises and the optimizer acting upon this information causes the damping factor to increase to quench these oscillations. The oscillations therefore decay and once again the damping factor drops to zero, causing a further burst of oscillations, this process being similar to that of squegging in an incorrectly biased oscillator.

Owing to limitations of the amplifiers and multiplier in the adaptive system, infinite damping for frequencies in excess of $\omega_0/\sqrt{2}$ cannot be obtained, since distortion sets in. Experimental results are also shown in Fig. 8, discrepancies for $\omega < \omega_0/\sqrt{2}$ being due to the continuous squegging in the system, which gives a finite mean damping factor.

A disadvantage of this adaptive system when subjected to a

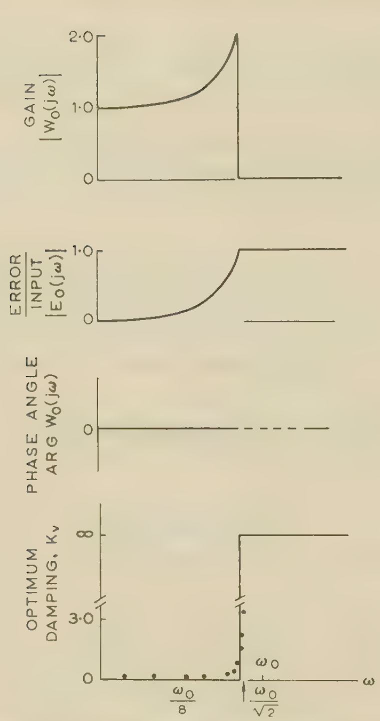


Fig. 8.—Optimum linear system characteristics for a sinusoidal input.

● Experimental results.

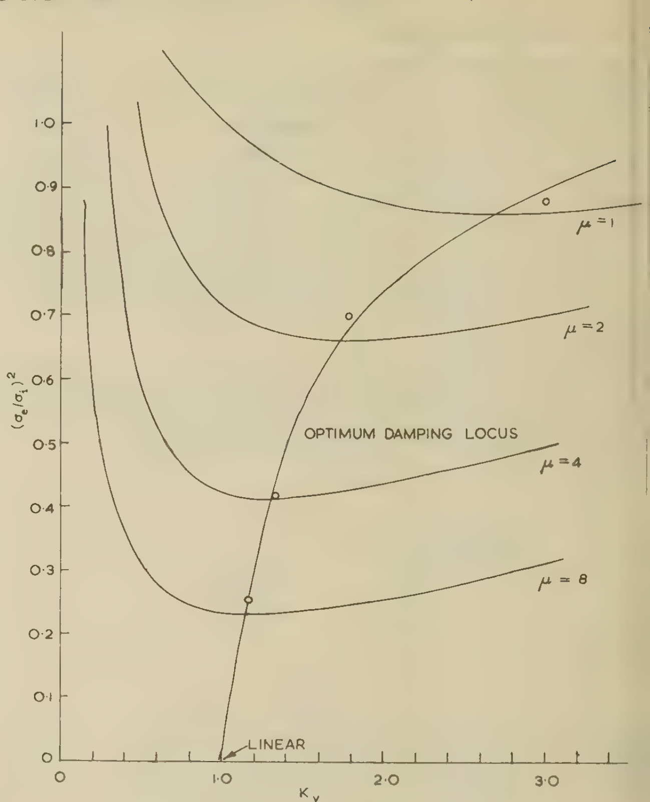


Fig. 9.—Linear system with random input.

○ Experimental minima.

repetitive input arises when the repetition frequency is very near that of the sampling frequency and its harmonics. Under these conditions large amplitude variations in parameter setting of beat frequency are observed.

(4.3) Random-Signal Performance

(4.3.1) Linear System.

With a knowledge of the input spectrum and the control-system transfer function the evaluation of error and output powers is relatively simple for the linear case.[10]

The transfer function of the linear system is of second order and its error ratio is given by

$$\mathrm{E}(p) = \frac{T^2p^2 + K_vTp}{T^2p^2 + K_vTp + 1}$$

and the input random signal is taken to have a simple spectrum of the form

$$\Phi_i(p) = \frac{\Phi_i(0)}{1 + \mu Tp}$$

Thus the mean squared error is

$$\sigma_e^2 = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} |\mathrm{E}(p)\Phi_i(p)|^2 dp$$

Substituting for $\mathrm{E}(p)$ leads to the result

$$\left(\frac{\sigma_e}{\sigma_i}\right)^2 = \mathrm{U}(\mu, K_v) = \frac{1}{K_v}\frac{K_v + \mu + \mu K_v^2}{1 + \mu K_v + \mu^2} \quad . \quad . \quad (1)$$

where

$$\sigma_i^2 = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left|\frac{\Phi_i(0)}{1 + \mu Tp}\right|^2 dp$$

is the mean squared input signal.

For optimum performance on the basis of the minimum mean-squared-error criterion, the expression for σ_e^2 in eqn. (1) must b[e] differentiated with respect to the variable parameter K_v, an[d] for a minimum,

$$\frac{\partial \mathrm{U}(\mu, K_v)}{\partial K_v} = 0$$

resulting in a quadratic in K_v^*:

$$\mu^2 K_v^{*2} - 2\mu K_v^* - (1 + \mu^2) = 0$$

whence

$$K_v^* = \frac{1}{\mu}[1 + \sqrt{(2 + \mu^2)}]$$

since $K_v < 0$ is not allowed from stability considerations.

Thus, for any value of bandwidth ratio, μ, there exists a fini[te] value of K_v giving minimum mean squared error. The variatio[n] of error/input power ratio as a function of damping factor K_v is shown in Fig. 9 for four values of μ.

It is to be noted here that for $K_v < K_v^*$ the rate of change [of] error power with damping factor for a given input is in gener[al] large, whereas for $K_v > K_v^*$ the slope is small. This inheren[t] asymmetry of the loci leads to the fundamental characteristic [of] the adaptive system, discussed in Section 2.1.

For misadjustments in K_v below K_v^*, rapid correction resul[ts] since the speed of response of the adaptive loop is a dire[ct] function of slope $\partial\mathrm{U}/\partial K_v$ of the loci. Conversely, for displac[e]ments in K_v above K_v^*, $\partial\mathrm{U}/\partial K_v$ is small, and a long time [is] necessary for the system to regain equilibrium.

(4.3.2) Torque-Limited System.

Where a control system subjected to a random input sign[al] involves non-linear elements within the closed loop, comple[te] analysis can be very tedious. An approximate analysis [is] possible by applying the quasi-linearizing technique of Booton[.] In this, the non-linear functions can be represented by equivale[nt]

linear relationships, whose magnitudes are dependent on the signal amplitudes applied to them. Booton has shown that the 'equivalent gain' of a non-linear function f(x) is given by

$$K_{eq} = \frac{1}{\sigma^2}\int_{-\infty}^{\infty} x\mathrm{f}(x)\mathrm{p}(x)dx \quad . \quad . \quad . \quad . \quad (2)$$

where σ^2 is the input power to the non-linear element and p(x) is the amplitude probability-density distribution of this signal. The random input signal is assumed to have a Gaussian amplitude distribution, and a further basic assumption is that this distribution is maintained throughout the system. It has been shown[7] that the low-pass filtering action of the servo motor tends to restore the amplitude-limited output signal from the torque limiter to the original Gaussian form.

The equivalent-gain technique enables the steady-state operating conditions to be evaluated graphically. Representing the gain of the torque limiter by the equivalent gain K_t, the transfer function of the quasi-linear system is

$$\mathrm{W}_1(p, \sigma_i) = \frac{K_t}{T^2p^2 + K_vK_tTp + K_t}$$

When the non-linear function is replaced by its equivalent gain K_t, two relationships can be obtained between K_t and the control-signal power, σ_c^2, applied to it. The first relationship is derived from quasi-linear theory and involves the control-system and input-signal parameters, and the second from eqn. (2).

Thus, with a knowledge of the control ratio

$$\mathrm{C}_1(p, \sigma_i) = \frac{x_c}{x_i} = \frac{T^2p^2}{T^2p^2 + K_vK_tTp + K_t}$$

and the input-signal spectrum $\Phi_i(p)$, application of Parseval's theorem gives the control power:

$$\overline{x_c^2} = \sigma_c^2 = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} |\mathrm{C}_1(p, \sigma_i)\Phi_i(p)|^2 dp$$

The input-signal power spectrum is taken to be of the simple form

$$\mathrm{G}_i(\omega) = \frac{\Phi_i^2(0)}{1 + \mu^2\omega^2T^2}$$

which has a 6 dB per octave high-frequency fall-off and a half-power angular frequency of $1/\mu T$. Substituting for σ_c^2 therefore yields

$$\sigma_c^2 = \frac{\sigma_i^2}{K_v}\,\frac{K_v + \mu}{1 + \mu K_vK_t + \mu^2K_t} \quad . \quad . \quad . \quad . \quad (3)$$

The torque-limiting non-linearity is taken to have unity slope in the linear regime $-h \leqslant x_c \leqslant h$ and zero slope outside this, resulting in an equivalent gain given by

$$K_t = \operatorname{erf}\frac{h}{\sigma_c\sqrt{2}} \qquad \text{[from eqn. (2) and Reference 6]}$$

$$= \operatorname{erf}\left(\frac{1}{\sigma_{iN}}\,\frac{\sigma_i}{\sigma_c}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where $\sigma_{iN} = \sigma_i\sqrt{2}/h$ is the normalized r.m.s. input voltage to the system. Eqns. (3) and (4) must hold simultaneously. For any given value of μ, plotting the two expressions as functions of the normalized input power σ_{iN}^2 and the variable parameter K_v will give the operating value of K_t. An example is shown in Figs. 10 and 11 for the case of $\mu = 8$. Fig. 10 is plotted from eqn. (3) and Fig. 11 from eqn. (4).

The operating value of K_t being known, the mean squared error in the system for any input power can readily be derived, since

$$\overline{x_e^2} = \sigma_e^2 = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} |[1 - \mathrm{W}_1(p)]\Phi_i(p)|^2 dp$$

$$= \frac{\sigma_i^2}{K_v}\,\frac{K_v + \mu + \mu K_tK_v^2}{1 + \mu K_vK_t + \mu^2K_t}$$

The variation of error power with input power and damping factor, K_v, is shown in Fig. 12 for the particular case of $\mu = 8$. In general, there will exist some value of K_v for each input power for which the mean squared error is a minimum.

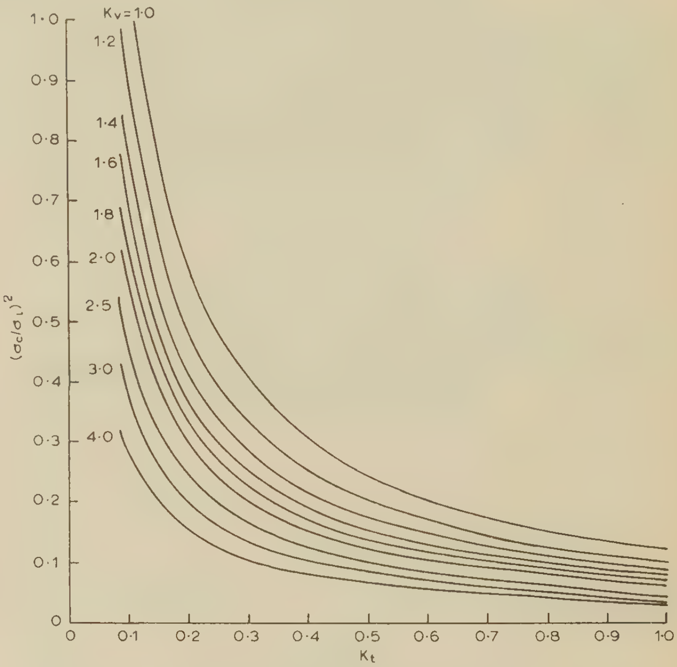


Fig. 10.—Self-adaptive system: variation of control power with damping ($\mu = 8$).

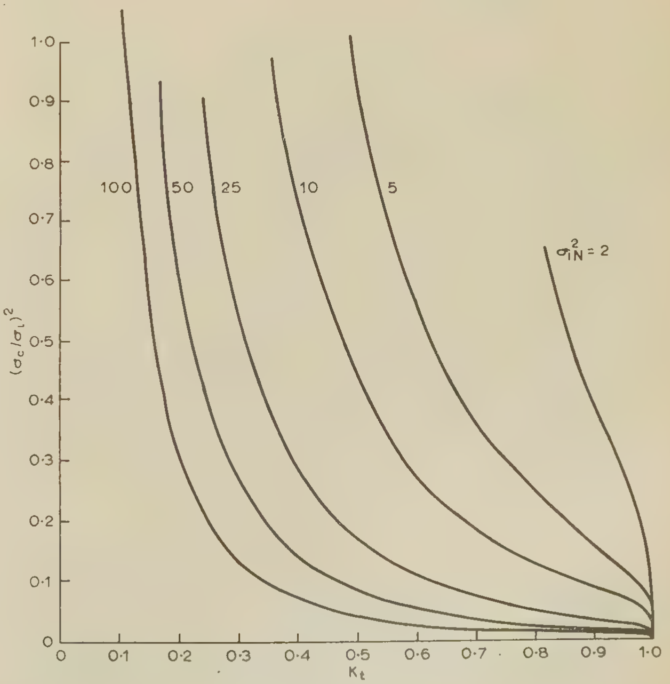


Fig. 11.—Derivation of the closed-loop response.

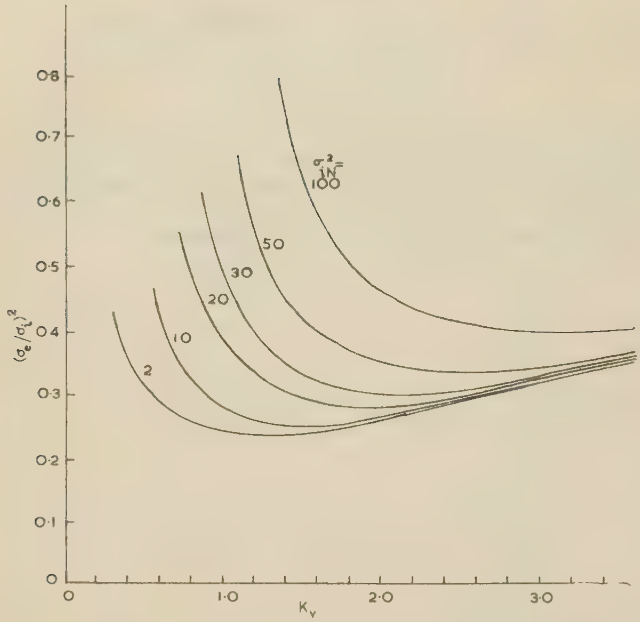


Fig. 12.—Self-adaptive system: variation of error power with input power and damping ($\mu = 8$).

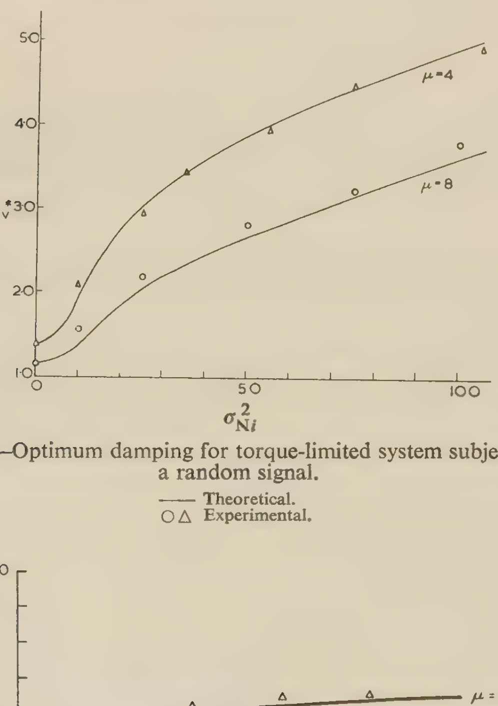


Fig. 13.—Optimum damping for torque-limited system subjected to a random signal.

—— Theoretical.
○△ Experimental.

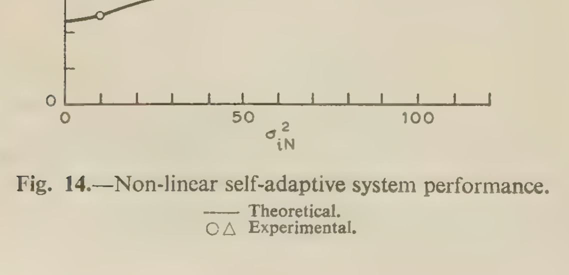


Fig. 14.—Non-linear self-adaptive system performance.

—— Theoretical.
○△ Experimental.

Fig. 13 shows theoretical and experimental results obtained with the self-optimizing system for two values of bandwidth ratio. The experimental value of optimum damping ratio, K_v^*, is in fact a time average or mean value, and has been designated the mean optimum damping factor. The variation of error power with input power for these two cases is shown in Fig. 14 and, as in Fig. 13, experimental results are in very good agreement with theory.

(5) CONCLUSIONS

A simple technique has been described for the automatic adjustment of one parameter of a system. It has been shown how a particular servo mechanism may be modified so that the damping is varied to give minimum system error, giving appreciable improvement where a wide range of input signals may be applied.

Analogue-computer analysis shows that the steady-state conditions lie near to the theoretical optimum and indicate that the performance of the system is equal to or better than previous non-linear compensation systems.

Further work will investigate the possibility of extending the self-adjusting facility to two or more parameters of the system, and will consider in more detail the stability of the optimizing feedback loops.

(6) ACKNOWLEDGMENT

The authors wish to acknowledge the help given by Mr. P. H. Hammond of the National Physical Laboratory, with whom many useful discussions have been held, and who provided facilities for initial experimental work.

(7) REFERENCES

(1) West, J. C. Douce, J. L., and Naylor, R.: 'The Effect of the Addition of Some Non-Linear Elements on the Transient Performance of a Simple R.P.C. System possessing Torque Limitation', *Proceedings I.E.E.*, Paper No. 1549 M, August, 1953 (**101**, Part II, p. 156).

(2) Douce, J. L., and King, R. E.: 'The Effect of an Additional Non-Linearity on the Performance of Torque-Limited Control Systems subjected to Random Inputs', *ibid.*, Monograph No. 1, 361 M, February, 1960 (**107** C, p. 190).

(3) West, J. C., and Nikiforuk, P. N.: 'The Frequency Response of a Servo-Mechanism designed for Optimum Transient Response', *Transactions of the American I.E.E.*, 1956, **75**, Part II, p. 234.

(4) Draper, C. S., and Li, Y. T.: 'Principles of Optimalizing Control Systems' (A.S.M.E., New York, 1951).

(5) Williams, F. C.: 'The Use of Velocity-Squared Damping for High-Speed Overloaded Servos', T.R.E. Report, 1942.

(6) Booton, R. C.: 'Non-Linear Control Systems with Statistical Inputs', Report No. 61, Dynamic Analysis and Control Laboratory (M.I.T., Cambridge, Mass.).

(7) Barrett, J. F., and Coales, J. F.: 'An Introduction to the Analysis of Non-Linear Control Systems with Random Inputs', *Proceedings I.E.E.*, Monograph No. 154 M, November, 1955 (**103** C, p. 190).

(8) Stromer, P. R.: 'Adaptive or Self-Optimizing Control Systems: a Bibliography', *Transactions of the Institute of Radio Engineers*, 1959, AC–4, p. 65.

(9) King, R. E.: Ph.D. Thesis, Queen's University of Belfast, June, 1960.

(10) James, H. M., Nichols, N. B., and Phillips, R. J.: 'Theory of Servo-Mechanisms' (McGraw-Hill, 1947).

(11) Taylor, W. R.: 'An Experimental Control System with Continuous Automatic Optimization', I.F.A.C. Conference, Moscow, 1960.

621.317.39

The Institution of Electrical Engineers
Paper No. 3590 M
July 1961

THE ELECTRICAL DETERMINATION OF MOISTURE IN PAPER

By T. S. McLEOD, M.A., Associate Member, and A. E. YALLUP, A.M.I.Mech.E., Graduate.

(*The paper was first received* 19*th January, and in revised form* 7*th March,* 1961.)

SUMMARY

Absorption of moisture causes paper to change its electrical resistance and permittivity. The second effect is approximately linear and little affected by changes in furnish. It is most conveniently measured by means of a fringing capacitor. Instruments for use statically and on a moving web are described. Theories of the water-cellulose bond are discussed.

(1) INTRODUCTION

(1.1) Importance of Paper Moisture Content

When paper is exposed to the atmosphere it rapidly reaches an equilibrium moisture content which depends upon the relative humidity and the temperature. There is a marked hysteresis at moderate humidities, a moist paper coming to equilibrium at a higher moisture content than a dry one when they are brought into the same atmosphere, as shown in Fig. 1. The individual

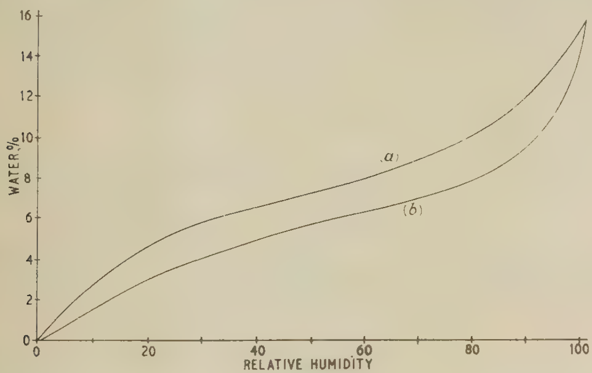


Fig. 1.—Relative humidity and equilibrium paper moisture content
(*a*) Desorption.
(*b*) Absorption.

cellulose fibres swell as they take up moisture and all the physical properties of the paper change. Dry paper is brittle and difficult to handle, and moist paper is weak. The optimum moisture content for most purposes lies between 5 and 10%, but paper can be made and used with between 2 and 14% by weight of water.

Paper is usually reeled up as it is made, and when it is cut up into sheets they are frequently stored in stacks. In either case the edges will come into equilibrium with the ambient atmosphere but the centre will remain as it was made. This produces crinkling at the edges and when the paper is unreeled for use it will be unstable until it has reached equilibrium with the atmosphere. For this reason it is important to make it with the correct moisture content, and to use it in an atmosphere compatible with this content. It can be seen from Fig. 1 that optimum water contents are compatible with most working atmospheres.

(1.2) Points at Which Measurement is Required

The paper maker's first requirement is to measure the moisture content of paper as it is made, and this should be done just before it is reeled up. This requires a measurement upon a moving web. A second requirement is a static measurement which can be applied to sheets ready for conversion or printing to ensure that they are stable.

As paper is sold by weight, it is also useful to be able to check the amount of water that is being invoiced.

(2) MOISTURE AND RESISTANCE

The ohmic resistance of paper depends upon its moisture content. Fig. 2 is a typical graph which shows the limitations of instruments based upon this effect.

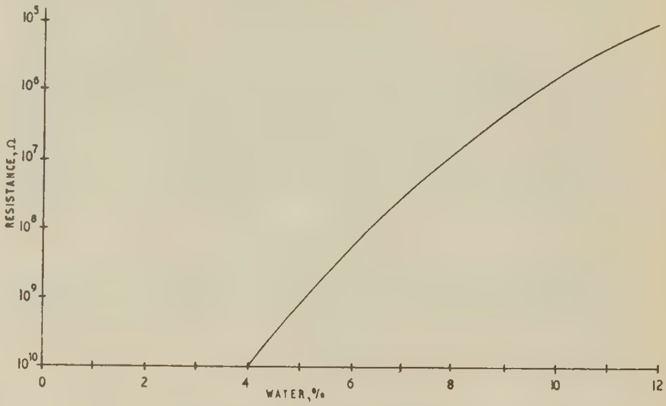


Fig. 2.—Paper moisture content and resistance.

At quite common moisture contents resistances are too high to be easily measured and at the other end of the range they are very low. Another difficulty is that small changes in the materials from which the paper is made can have a marked effect upon its resistance. Nevertheless, the resistances are reasonable over the part of the scale on which most paper makers wish to work and some mills have made excellent use of resistance instruments. In many cases their practice is to disregard the original calibration of the instruments, to get the paper running in the correct condition as determined by gravimetric moisture checks, and to instruct the machine men to keep the moisture meter on the reading which it then shows.

(3) MOISTURE AND PERMITTIVITY

(3.1) Values of the Permittivity of Paper

The permittivity of most papers is between 1·5 and 3 when dry and increases fairly linearly between 2 and 14% moisture content. Fig. 3 shows results for four papers. In three cases the rate of increase of permittivity with water content is proportional to the amount by which the initial permittivity exceeds unity. This is to be expected since the initial value depends

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. McLeod is with the Plessey Co., Ltd. Mr. Yallup was formerly with the British Paper and Board Industry Research Association and is now at the British-American Tobacco Company Research Establishment.

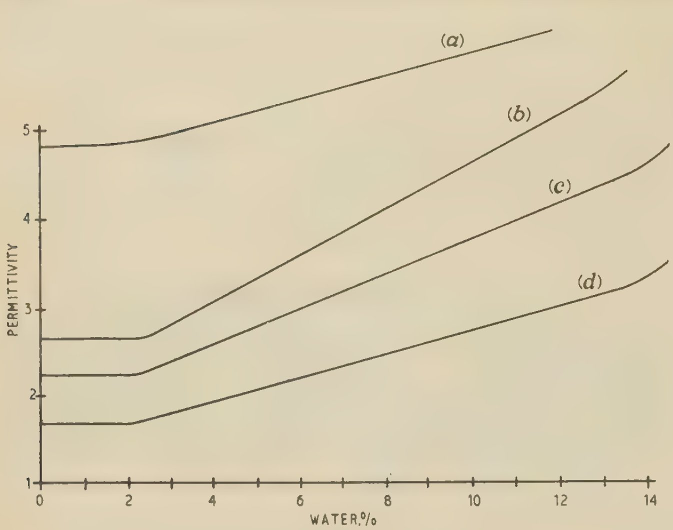


Fig. 3.—Paper moisture content and permittivity.
(a) TiO_2 loaded.
(b) Drawing.
(c) Greaseproof.
(d) Mechanical printing.

upon the relative amounts of fibre and air in the paper, and the amount of water for a certain percentage increase of weight is proportional to the initial weight of fibre. The fourth paper contained an additive of high permittivity. The initial value is high but the slope is low. This is because the initial high value is due, not to dense packing of the fibres, but to the additive which contributes little weight and takes up no water but makes a substantial contribution to the permittivity. This effect is found in papers loaded with carbon, titanium dioxide and certain plasticizers.

(3.2) Q-Factors

Fig. 4 shows the Q-factors at various frequencies of capacitors with paper dielectrics of high moisture content. It will be seen

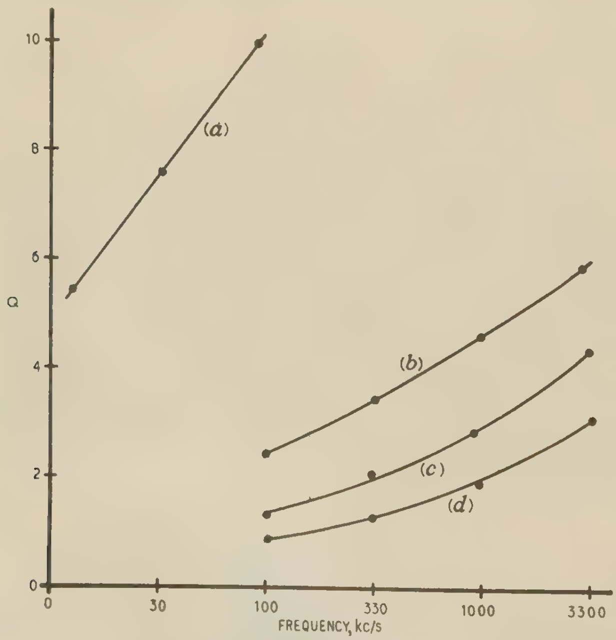


Fig. 4.—Q-factors for a capacitor with paper dielectric.
(a) Drawing, 9% water.
(b) Newsprint, 17% water.
(c) Mechanical printing, 15% water.
(d) Drawing, 14% water.

that, in general, frequencies above 1 Mc/s must be used if water contents as high as 14% are to be measured by capacitive instruments. At lower frequencies, the ohmic conductance obscures changes in permittivity. The Q-factor rises rapidly as the water content falls, and at 9% a frequency as low as 10 kc/s is usable.

(3.3) Characteristics of Capacitance-Type Instruments

It is clear from Fig. 3 that a capacitance-measuring instrument has distinct merits. The change in permittivity with paper moisture content is approximately linear over the whole of the required range. Measurements may be made right down to 2% and up to 14% provided that a frequency greater than 1 Mc/s is used. The slope of the curve depends upon the amount of fibre in the paper and additives only shift the zero. An instrument may therefore be calibrated with sensitivity in terms of paper basis weight, and a correction may be made for additives in terms of a furnish factor which represents a shift in the zero of the scale.

(4) CONTACTLESS MOISTURE MEASUREMENT

In the manufacture of paper, an equilibrium must be established between the moisture in the paper and that in the air immediately above it. Several instruments have made use of this principle. They consist of hygrometric elements which are suspended just above the web. They are, unfortunately, liable to be upset by draughts and by changes in ambient temperature. Very sharp humidity gradients are frequently established above a paper web, and so their placing is critical. They have the limitations common to all known hygrometers, as recently analysed in a National Physical Laboratory pamphlet.[1] A number of instruments of this type are in use but they are not widely accepted in the industry.

(5) CAPACITOR DESIGN

(5.1) Parallel-Plate Capacitor

The simplest system is to guide the paper between parallel plates. The capacitance is then given by

$$C = \frac{A}{4\pi[d - t(1 - 1/\epsilon)]}$$

where A = plate area, d = plate separation, ϵ = permittivity and t = thickness of paper.

The equation indicates the disadvantages of this system. Since the moving paper must be threaded through the plates, d must be large compared with t. This makes the variation of C with ϵ non-linear, and less sensitive for large values of ϵ than for small ones. The good point is that the position of the paper in the gap does not matter, nor need it be in contact with either plate.

(5.2) Fringing Capacitor

The fringing capacitor consists of flat parallel electrodes lying in the same plane. The paper dielectric is laid in contact with the electrode faces and produces a change in capacitance proportional to its permittivity. The effect of thickness upon capacitance defies exact calculation, but its general form can be determined experimentally. The relation is at first linear, and then tails away. The field pattern deduced from the measurements is sketched in Fig. 5. At 1/40th of the electrode separation the field has fallen by 10% and at 1/20th, by 30%.

An important advantage of this type of capacitor is that the effect upon its capacitance of a paper dielectric depends principally upon the amount of fibre in a given area, and only to a second order upon the thickness of the sheet. Consequently,

Fig. 5.—Fringing capacitor field distribution.

he effect of a particular type of paper is proportional to its veight per unit area, irrespective of its thickness, and this applies qually to moisture contained in the paper.

(5.3) Capacitor Design

The linear relation between capacitance and permittivity tells heavily in favour of the fringing capacitor, but its crucial advantage is that it can be laid on top of a paper web and swung out of position when necessary. No paper threading is required.

The size of the electrodes is determined by the strength of paper. They must be large enough to support the weight of the electronic components above them, they must be thick enough to withstand rough handling without distortion, and they must be far enough apart for the electric field to penetrate effectively into the paper and give an indication of overall moisture content and not just of that of the surface layer.

Electrodes measuring $\frac{1}{2}$in $\times$ $\frac{1}{4}$in $\times$ 6in meet these requirements, although none of the dimensions are critical. To deal with paper up to 0·025in thick, a separation of at least $\frac{1}{8}$in is required. This gives a capacitance change of only 0·5pF for a change of 12% in the water content of paper 0·003in thick, which puts severe requirements on the electronic circuits.

(6) ELECTRONIC CIRCUIT DESIGN

(6.1) The Blumlein Bridge

An excellent circuit for the measurement of small capacitances was described by Blumlein 25 years ago.[2] Referring to Fig. 6

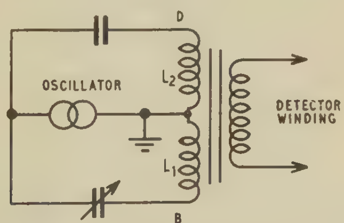


Fig. 6.—Blumlein's bridge circuit.

the standard and the measured capacitances balance the inductive arms L_1 and L_2, which are wound in opposite directions on the same core. At balance the currents in the two arms are equal and the self-inductance of each winding is cancelled, apart from leakage, by its mutual inductance with the other. Consequently at balance B and D are both at earth potential and there is no flux in the transformer. A small change in one capacitance produces a proportional flux and gives an output in the detector winding directly proportional to the change of capacitance.

(6.2) Transistorized Version

A development of Blumlein's circuit is shown in Fig. 7. The bridge and output windings are on a high-permeability toroid so that leakage is negligible. A phase-sensitive detector is used so that capacitance changes are measured but changes in resistance are not important. For convenience in construction, one side of the capacitor is earthed, thus sacrificing one merit of the original circuit. As the only active element is a simple

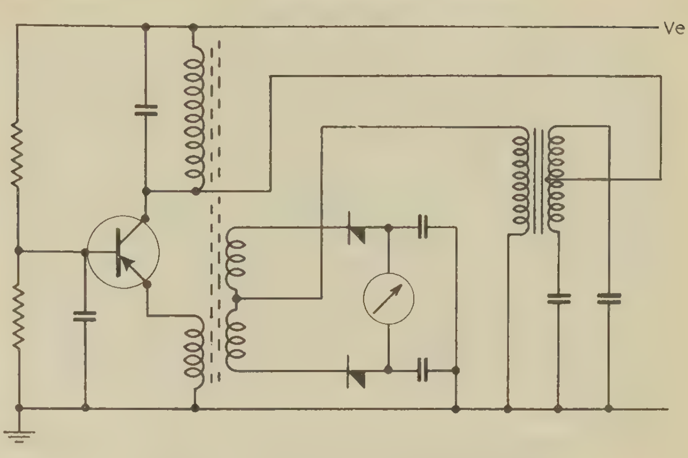


Fig. 7.—Complete phase-sensitive bridge.

transistor oscillator, the entire circuit can be mounted upon the electrode structure and the output can be taken as a d.c. signal to a remote indicator. Section 13 gives a full analysis and shows that with a standing electrode capacitance of 25pF an output of 96mV/pF is available.

(7) INSTRUMENT DESIGN

(7.1) Dynamic Instrument

The first requirement is that the capacitor electrodes should ride evenly upon the paper web and that the output should be indicated remotely. In Section 5 it was shown that measurements must be made at a frequency greater than 1 Mc/s and that the electrode structure should be several inches in length and breadth. The development of transistor circuits has made it possible to build an r.f. bridge on the lines described in Section 6 and to mount it directly upon the electrodes. It is then possible to take a d.c. output back by the multicore cable carrying power supplies and to put the meter or recorder at any convenient point. Limitations of germanium-transistor performance restrict these devices to a maximum temperature of 60°C, which is seldom exceeded a few inches above the paper web, but silicon transistors can be used up to 100°C.

In use, instruments may be mounted either above or below the moving web. One counter-weight system ensures a light but firm pressure between the electrodes and the paper. Another maintains evenness of pressure along the length of the electrodes. Stops limit rotation of the head to a few degrees each way so that it can follow the movement of the paper but cannot be turned over by sudden whipping or lurching.

(7.2) Static Instrument

The general principles governing static measurement are the same as those for dynamic measurement. It is necessary to use concentric or crossed electrodes rather than parallel ones since the permittivity of paper is directional, being greatest perpendicular to the direction in which the fibres tend to lie. This produces apparent differences up to $\pm 1\%$ in indicated moisture content as an instrument with parallel electrodes is rotated. This effect is not seen in the dynamic instrument, as its electrodes always lie parallel to the web. The paper is supported upon a block of foamed plastic since any other material is likely to have a permittivity as high as that of paper and to increase the standing capacitance in the measurement. The pressure of the head upon the paper must be constant and this is achieved by allowing the head to rest supported upon the paper while a measurement is being made.

(7.3) Choice of Insulators

It is clear that the insulators used to support the live electrodes in these instruments must stand severe conditions of heat and humidity without change of permittivity, either directly or as a result of the absorption of moisture. Glass and porcelain absorb surface films of moisture. P.T.F.E. is excellent cold but lacks mechanical strength at high temperatures. Loaded p.t.f.e. absorbs moisture. Most solid plastics change their permittivities with temperature, and laminar structures absorb moisture. Protective films of silicone varnish or Araldite slow this process but do not eliminate it. The most satisfactory material so far discovered has been natural amber, although even with this a very slight surface film of moisture is formed at high humidities and it is necessary to make the surface leakage paths as long as possible.

(8) CALIBRATION FOR DIFFERENT TYPES OF PAPER

(8.1) Basis Weight

The theory of the dependence of permittivity upon the moisture content of paper is discussed in Section 10; it is incomplete and uncertain. Practical experience happily shows the relation to follow the simple curves of Fig. 3.

In each case the graph follows an approximately straight line between 2 and 14%. Comparison of this with the capacitor performance described in Section 5 shows that the amplifier sensitivity required depends only upon the basis weight of the paper. Fig. 8 shows the gains required, taking a basis weight of 40 g/m^2 as unity. This makes calibration a straightforward matter.

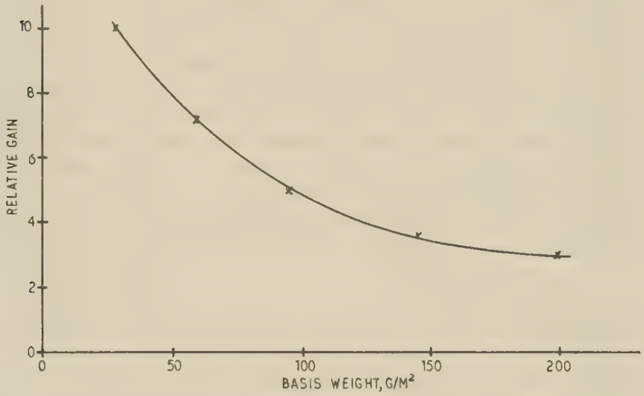


Fig. 8.—Variation of bridge sensitivity with paper basis weight.

(8.2) Furnish Factor

The curves of Fig. 3 intercept the zero axis at different points. This is because papers of the same basis weight have different permittivities when dry. This may be due to the nature of the fibre or to additives. In particular, papers containing carbon, titanium dioxide or plasticizers have high permittivities. Table 1 shows the points at which the projected straight-line part of the curves for a variety of papers intercept the zero axis. Strictly, a correction, known as the furnish factor, should be determined for each type of paper by a gravimetric check. In practice, it is usually adequate to take one of the values given in the Table as the error is unlikely to be greater than $\pm\frac{1}{2}\%$.

Table 1

FURNISH FACTORS FOR COMMON TYPES OF PAPER

Cotton rag (blotters, etc.)	0·5–1·5
Newsprint	1·5–2·0
Mechanical printing	1·5–2·0
Sulphite esparto	1·8–2·2
Foudrinier and M.G. kraft	2·0–2·5
M.G. sulphites	2·0–2·5
Sisal and manila rope	3·0–4·0
Papers with TiO_2, carbon or plasticizer	> 4

The provision of a control which shifts the zero of the amplifier output by a certain percentage of indicated moisture is straightforward.

(9) OPERATIONAL EXPERIENCE

(9.1) Mechanical Performance

The mechanical problems of riding the instrument upon a moving paper web have proved less severe than might have been expected. Even the lighest papers have proved strong enough to support it and there has been no tendency to bounce. An occasional violent movement of the paper throws the head into the air, but it alights safely so long as it is prevented from rotating and coming down on its side. Stainless-steel runners have shown little wear in hundreds of hours of use, and slight pitting caused by electrostatic discharge from very dry paper has not affected calibration.

(9.2) Consistency of Results

The recordings of Figs. 9 and 10 are typical. A dry paper containing about 5% of moisture shows fluctuations of about $\pm\frac{1}{2}\%$. It was at first thought that these fluctuations were instrumental, but when an instrument was mounted close to another type working at a different frequency they followed one another closely. This showed that the fluctuations were genuine. Papers with higher moisture contents are more erratic. When the mean value is about 10% slight changes in machine speed, steam pressure or ambient conditions produce a severe effect and a variation as small as $\pm 1\%$ is seldom maintained for long. Most paper makers have preferred to have instruments with time-constants of 2 or 3 sec so that rapid fluctuations are averaged out.

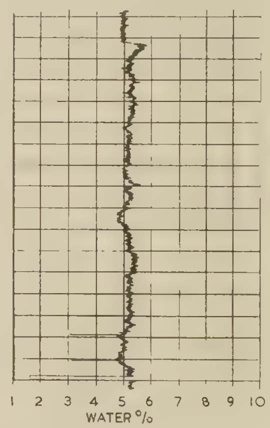


Fig. 9.—Paper running dry, 1 hour in detail.

(9.3) Accuracy

Fig. 11 shows a number of spot gravimetric checks taken on a web. Unfortunately the process of cutting a sample from a moving web, weighing it, drying it out and weighing again, is fraught with difficulties. The best techniques show an average scatter of about 0·5% in samples taken simultaneously, and it can be shown that if the mean of three is taken the probability is that in 95% of cases the result will be within $\pm 0{\cdot}5\%$ of the true value. This is shown in Fig. 11 and it is clear that the

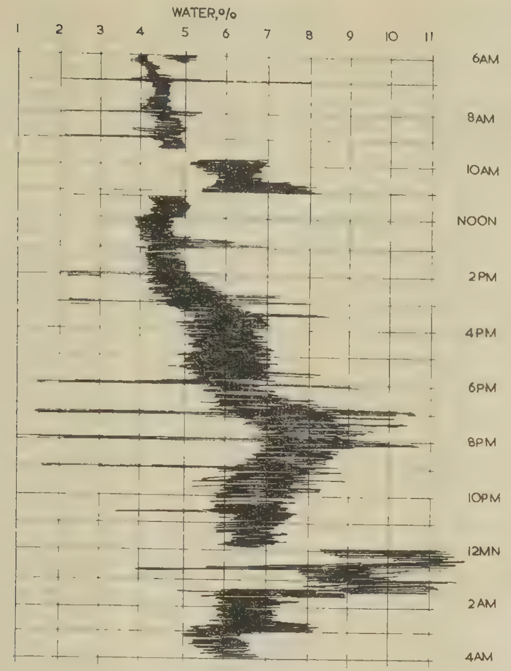


Fig. 10.—A 24-hour recording in a mill making 0·025 in chipboard.

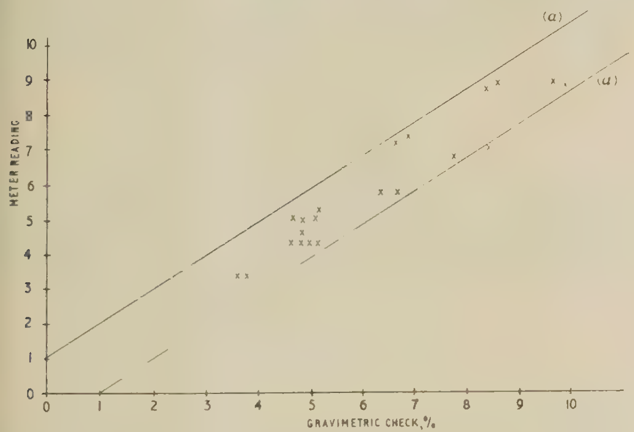


Fig. 11.—Comparison of meter recording and gravimetric moisture determination.

(*a*)–(*a*) 95% limit lines.

accuracy of the meter is at least as great as that of the best method of checking it.

Errors in basis-weight setting produce errors in indicated moisture content. $\pm 5\%$ in basis weight will produce an indicated error of $\pm 0{\cdot}2\%$ in moisture in a paper of 40 g/m^2 and 5% water, increasing to $\pm 0{\cdot}4\%$ for 10% water. Owing to the shape of the field the error for thicker papers is smaller, and for 250g/m^2 paper running $\pm 5\%$ of correct weight the error will only be half these figures.

(10) THEORETICAL INTERPRETATION OF RESULTS

(10.1) Nature of the Fibre-Water Bond

This subject has been discussed by many authors and there is little agreement between them. Wood[3] gives an excellent summary and Muus[4] outlines the various mathematical treatments. The experimental facts are as follows. Paper rapidly takes up or loses moisture, reaching an equilibrium which depends upon the relative humidity and temperature of the air surrounding it. The curves (Fig. 1) show hysteresis, i.e. dry paper reaches equilibrium at a lower moisture content than moist paper when they are brought into an atmosphere with a humidity corresponding to an intermediate value. As the moisture content increases the binding energy falls, so that less heat per gramme is required to remove the moisture at higher moisture contents. Conversely, more heat is released when moisture is absorbed by dry papers than by wet. The first 2% increase of weight of the paper produces little change in its permittivity: after that the curve rises with increasing steepness (Fig. 3). During this early stage the fibres show little change in size, but as the regain increases each fibre swells laterally.

Most of the theories postulate that there are areas in a fibre structure which take up water forming some type of chemical bond. This water becomes part of the structure of the fibre and consequently has a negligible effect on its permittivity or resistivity. Its binding energy is large, between 200 and 300 cal/g. This accounts for the fact that the first 2% of water is taken up within a few seconds, even in air of very low humidity, and is only removed by prolonged heating. Some authors have been puzzled by the figure of 2% which indicates that one water molecule is taken up for every three cellulose units.

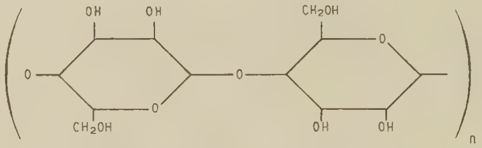


Fig. 12.—Cellulose structure.

Hailwood and Horrobin[5] suggest that the explanation is to be found in the partially crystalline nature of cellulose revealed by X-ray analysis. If, in paper, about two-thirds of the cellulose is crystalline and one-third disorganized, then there is one molecule of water for one unit of non-crystalline cellulose. The crystalline proportion of cellulose is reasonably near this figure, so if we postulate that non-crystalline cellulose has units with free valencies and that each valency takes up one water molecule, the phenomenon is explained. As this 2% is effectively in chemical combination with the cellulose it has no appreciable effect upon either resistance or permittivity.

After the first 2% the absorption of water takes place with an approximately linear rise in permittivity, an approximately linear lateral swelling of the fibres and a logarithmic decrease in resistance. The binding energy is about 25 cal/g. This is readily explained by capillary absorption of water between the fibres. The linear increase in permittivity is similar to that in other water/solid mixtures. The logarithmic fall in resistance is again typical of these. The swelling can be interpreted as a simple pushing apart of the fibres.

(10.2) Permittivities

The effect of moisture upon the permittivity of paper has been investigated by Garton,[6] who considers two possibilities. Either the water may form discrete globules, or the molecules may be evenly distributed. If the dry paper has permittivity ϵ_1, the moist paper ϵ_2, and the water considered either as a cloud or as a gas ϵ, then for the first case we have

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{3(\epsilon_2 - \epsilon_1)}{(\epsilon_1 + 2)^2}$$

and for the second case,

$$\epsilon - 1 = \epsilon_2 - \epsilon_1$$

Garton shows that the second formula is in better agreement with his experimental results, which followed the linear relationship confirmed by the present writers. The linear relationship begins to break down when the water starts to form globules, the point at which paper rapidly loses strength.

(11) ACKNOWLEDGMENTS

The authors wish to express their indebtedness to Dr. N. R. Hood and Mr. W. E. Bennett of the British Paper and Board Industry Research Association for direction and help throughout this work; to Mr. C. M. Deavin of the Plessey Co., Ltd., for much work on the electronic circuits; and to Mr. P. Boyle and the late Mr. M. D. Holmes of the Baldwin Instrument Co., Ltd., for carrying out trials in paper mills which led to valuable improvements in the instruments described.

(12) REFERENCES

(1) 'Measurement of Humidity', National Physical Laboratory Pamphlet, 1958.
(2) Blumlein, A. D.: British Patent Specification No. 581161.
(3) Wood, H. J.: 'The Physics of Fibres', Institute of Physics Monograph, 1955, p. 22.
(4) Muus, L. T.: 'Dielectric Investigation of Cellulose with Special Consideration of the Cellophane-Water System', *Tansactions of the Danish Academy of Technical Sciences*, 1953, No. 4, p. 1.
(5) Hailwood, A. J., and Horrobin, S.: 'Absorption of Water by Polymers: Analysis in Terms of a Simple Model', *Transactions of the Faraday Society*, 1946, **42 B**, p. 84.
(6) Garton, C. G.: 'The Drying Process in Paper, as Determined by Electrical Methods', *Journal I.E.E.*, 1940, **86**, p. 369.

(13) APPENDIX

Calculation of Bridge Output

Consider a bridge with arms of inductance L, capacitances C and $C + dC$, and an applied voltage V (Fig. 6). Let the currents in the arms be i and $i + di$. The flux in the core will produce a voltage $V = j\omega Ldi$.

Then $$i = j\omega C(V + j\omega Ldi)$$

and $$i + di = j\omega(C + dC)(V - j\omega Ldi)$$

Neglecting $\omega^2 LdidC$,

$$\frac{di}{dC} = \frac{j\omega V}{1 - 2\omega^2 LC}$$

Choosing $2\omega^2 LC = \frac{1}{2}$ so as to be safely clear of resonance effects,

$$\frac{dV}{dC} = \frac{V}{4C}$$

When $V = 5\,\mathrm{V}$ and $C = 26\,\mathrm{pF}$, $dV/dC = 96\,\mathrm{mV/pF}$ for a detector winding of the same inductance as the bridge windings.

:.387.132.223.017: 621.376.5

The Institution of Electrical Engineers
Paper No. 3619 E
July 1961
©

MEASUREMENT OF ENERGY LOSSES IN A HYDROGEN-FILLED THYRATRON IN MODULATOR DUTY

By H. de B. Knight, M.Sc., F.Inst.P., Member, and J. LORD, D.L.C.

(*The paper was first received* 17*th November*, 1960, *and in revised form* 15*th March*, 1961.)

SUMMARY

The energy loss in a hydrogen thyratron during a current pulse may considered as made up of three components, occurring respectively the leading edge of the pulse, during the steady state and at the iling edge after the end of forward conduction. In the experi-nts described the three components were separated by pulse chop-ıg and the dissipation was measured by means of a calorimeter. .e conclusions were confirmed by tests with a saturable reactor in ies with the thyratron.

The leading-edge dissipation is predominantly that due to the charge of the anode-to-cathode capacitance, except that at low gas essures there is a pulse-front component which increases rapidly th decreasing pressure. The steady-state loss is proportional to the ean value of current. The trailing-edge loss increases directly with e rate of fall of pulse current, with the square of the inverse voltage d with its duration (up to the completion of deionization). The lse-front and trailing-edge components can be eliminated to a con-lerable extent by the use of a saturable reactor. The leading-edge sses in deuterium and in hydrogen are similar; but the steady-state d trailing-edge losses are 20–30% lower in deuterium.

The results are discussed in relation to the method of rating of drogen thyratrons for modulator duty.

(1) INTRODUCTION

The purpose of the work described is to study how the different mponents of energy dissipation in thyratrons filled with ydrogen vary with the operating conditions in a recurrent-pulse rcuit such as is illustrated in Fig. 1; such information provides

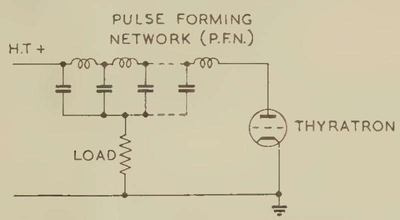


Fig. 1.—Pulse modulator circuit.

basis for the rating of thyratrons for modulator duty. The nergy dissipated consists of three main components, associated ith three phases of conduction of the current pulse, as shown in ig. 2; these are the leading-edge component, W_L, the steady-ate component, W_S and the trailing-edge component, W_T.

The leading-edge component may be considered in two parts, pulse-front component associated with the discharge of the ulse network and a capacitance component due to the discharge f the inter-electrode and stray capacitances between the anode nd the cathode. In the steady state the dissipation per pulse , to a first approximation, proportional to the pulse duration. he trailing-edge loss is associated with the reverse current after ne end of the forward-current pulse. A further possible source

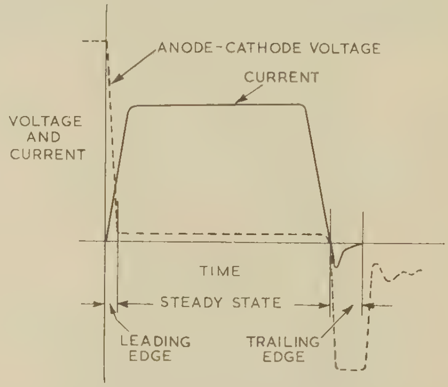


Fig. 2.—Typical waveforms of current through and voltage across thyratron.

Leading edge, steady state, and trailing edge indicate approximately the periods in which the losses W_L, W_S and W_T occur.

of energy loss is that due to electron current from the grid in the inter-pulse period; grid-emission measurements showed this to be negligible in the valves used for test. Measurements were made of the separate components of loss with hydrogen filling and also with deuterium filling.

(2) EXPERIMENTAL METHOD AND APPARATUS

In the tests the three components of loss were separated by a pulse-chopping technique, and the dissipation was measured by means of a calorimeter; the conclusions were confirmed by tests with a saturable reactor in series with the thyratron. The tests were made on two thyratrons similar to the type BT101 (which is rated at 25 kV, 500 A peak, 0·5 A mean); one valve was filled with hydrogen and the other with deuterium. The valves were operated at 25 kV, 260 A peak, except where it is otherwise stated.

(2.1) Pulse Chopping

Pulse chopping was carried out by means of circuits such as that shown in Fig. 3(*a*). In each case there are two thyratrons, V_1, V_2, in parallel. Conduction is initiated first in V_1 and then in V_2 after a given time interval t_1; a suitable firing voltage for V_2 is ensured by the inclusion of a resistor R [Fig. 3(*a*)] or a capacitor C (Fig. 4) in series with V_1. When V_2 is conducting, the voltage across it falls to the arc-drop value; this is too low for conduction through V_1 and the resistor, and current in that path falls to zero. Thus the energy W_L and a portion of W_S are dissipated in V_1 and the remainder of W_S plus W_T in V_2.

Ideally the current transfer would be instantaneous. In practice there is an overlap, as illustrated in Fig. 3(*b*) and (*c*), the duration of which depends on the time-constant of the loop containing the two thyratrons, and hence on the value of *R*. If the latter is small, current may still flow in V_1 up to the end of the total pulse [Fig. 3(*b*)]; in this case the proportion of W_S in this valve is difficult to determine, and if the current persists too lon there may also be some trailing-edge loss. If *R* is large

Written contributions on papers published without being read at meetings are vited for consideration with a view to publication.
The authors are with Associated Electrical Industries (Rugby), Ltd.

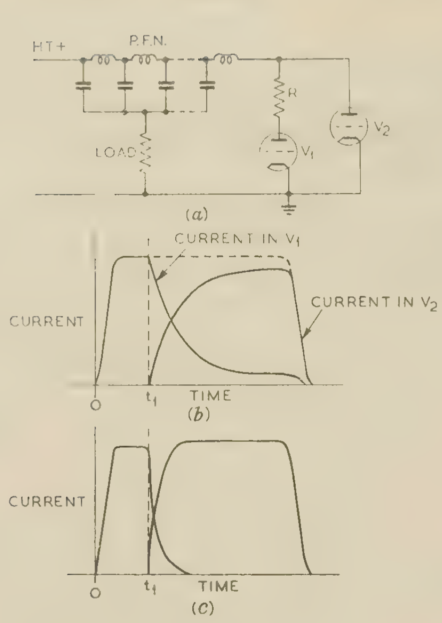


Fig. 3.—Pulse chopping, resistor method.

(a) Circuit diagram.
(b) Current in V_1 and V_2, $R = 5\,\Omega$.
(c) Current in V_1 and V_2, $R = 20\,\Omega$.

[Fig. 3(c)] the transfer is rapid but the firing voltage applied to V_2 may be so high as to entail some leading-edge loss. Also, the mismatch introduced by R leads to a delay in the application of inverse voltage, so that the conditions in V_2 differ from those met in practice. The resistor method is thus mainly useful for measurements of $W_L + W_S$ in V_1; some of the results given below were so obtained, with $R = 20\,\Omega$, and comparative tests made with the capacitor method confirm their accuracy.

The preferred method is that illustrated in Fig. 4. A capacitor, C, with a parallel discharging resistor is inserted in series with V_1. The voltage appearing across C due to the current flowing in V_1 reaches a value of about 1 000 V by the time V_2 is triggered; this is sufficient to ensure conduction in V_2 without introducing any appreciable leading-edge loss in it. The current decay in V_1 and the current rise in V_2 are of substantially uniform slope. This method is suitable for measurements of $W_L + W_S$ in V_1 and of $W_T + W_S$ in V_2.

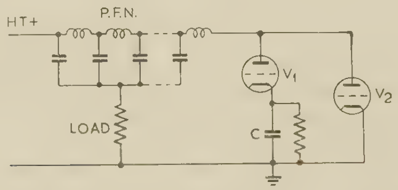


Fig. 4.—Pulse chopping, capacitor method.

The double-chopping circuit of Fig. 5 is an adaptation of the capacitor method. When V_2 is fired, the anode voltage on it is insufficient to cause leading-edge loss, and if V_3 is fired early enough to allow ample time (in our experiments $2\,\mu s$) for deionization in V_2 before the end of the total pulse, the trailing-edge loss in that valve is negligible. Hence the losses in V_2 may be assumed to consist only of the steady-state loss, W_S. The pulse length is measured at half amplitude: as the pulse shape is symmetrical, it is assumed that the actual losses are equivalent to what they would be with a constant-amplitude pulse of that length.

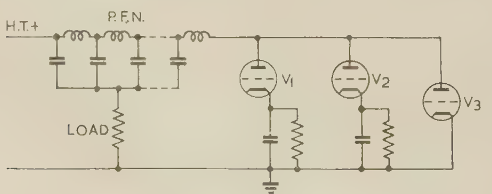


Fig. 5.—Double pulse chopping.

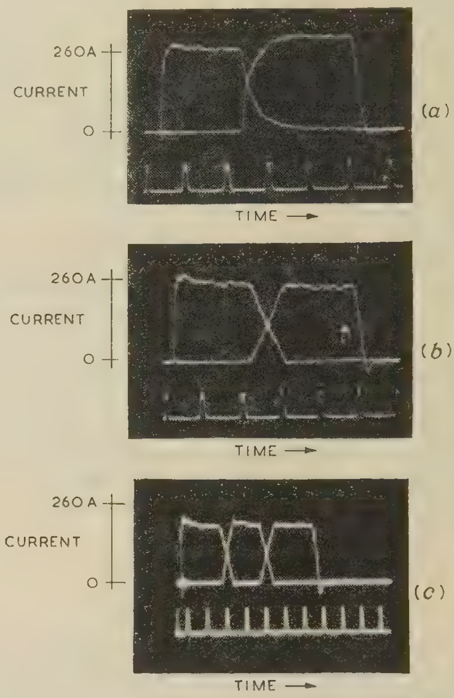


Fig. 6.—Pulse chopping, typical current oscillograms.

Calibration marks at $1\,\mu s$ intervals in each case.
(a) Resistor method.
(b) Capacitor method.
(c) Double-chopping method.

Fig. 6 shows the waveforms of the currents in the different valves in the experiments to be described, and in which loss measurements were made.

(2.2) Use of Saturable Reactor

A saturable reactor connected in series with the anode of the thyratron may be used to eliminate part of the leading-edge loss and practically all of the trailing-edge loss. In some experiments the use of the reactor was combined with pulse chopping, the reactor being connected in circuit as shown in Fig. 7.

At the start of the current pulse, the high inductance of the unsaturated reactor results in a delay period during which the current rises slowly and most of the applied voltage appears

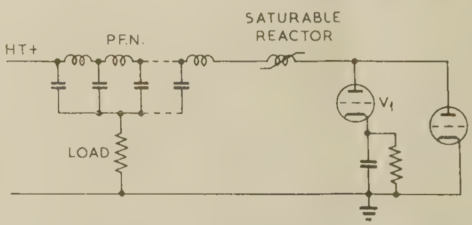


Fig. 7.—Saturable reactor in chopping circuit.

ross the reactor. During the delay, the voltage across the lve falls from the initial hold-off value to the arc-drop value hile the current is still very low; under these conditions the ading-edge dissipation, in so far as it is due to the relation tween pulse current and valve voltage, is negligible.[1] When turation occurs, the reactor has no appreciable effect on the haviour of the circuit.

At the end of the current pulse the reactor becomes unsaturated low values of forward or reverse current. Any inverse voltage the circuit now appears mostly across the reactor and this ndition continues until the reverse current reaches a value fficient to cause saturation. This effect has already been plied in power-rectification circuits to reduce the tendency to c-back,[2] and in modulator circuits to delay the appearance of verse voltage across the valve.[3] During the delay period some ionization takes place in the anode-grid space and as a result e energy dissipation associated with inverse voltage and reverse rrent is reduced.

The reactor designed for use in these tests is shown diagram-atically in Fig. 8. It consists of 14 toroidal cores of Permalloy

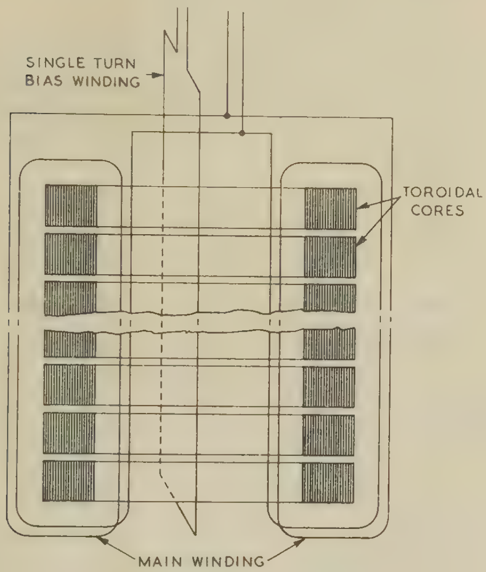


Fig. 8.—Experimental saturable reactor.

material, of $1\frac{1}{4}$ in inside diameter and $1\frac{7}{8}$ in outside diameter. wo windings, each consisting of 11 turns, are wound in parallel, ach winding covering one half of the cores; each turn was rmed of 3/20 s.w.g. copper wires in parallel. The reactor, hich was oil-immersed, was designed to work at 25 kV and to arry a current of 20 A r.m.s. in pulses of 500 A peak and 2 μs uration. The saturated and unsaturated inductances were 2 μH nd 3 mH respectively. The product of voltage and time quired for saturation, as measured, was about 0·012 Vs and e saturating current about 13 A. When it was connected in ries with the thyratron in a modulator circuit operating at 5 kV there was a delay of 0·5 μs on the leading edge of the rrent pulse, and in conditions leading to an inverse voltage of ·7 kV at the end of conduction the inverse voltage across the alve rose roughly exponentially to about 2 kV in 1 μs.

A single-turn biasing winding on the reactor was connected an auxiliary circuit from which current pulses of 1 μs duration ould be passed in synchronism with the pulses in the main rcuit. These could be timed to saturate the reactor just ahead the main pulse, so that the reactor was ineffective at the leading edge but effective at the trailing edge of the pulse. This gave an alternative method of separating the loss components.

(2.3) Separation of Losses

(2.3.1) Choice of Circuit.

It is not possible to isolate either the leading-edge loss, W_L, or the trailing-edge loss, W_T; each is accompanied by some portion of the steady-state loss, W_S, the amount of the latter depending on the duration of the pulse in the valve in question. It is necessary, therefore, first to estimate W_S before the values of the other components can be deduced.

Table 1 shows the combinations of loss components obtained with different circuits.

Table 1

Basic circuit	Valve	Reactor in circuit	Loss Components			
			W_{LP}	W_{LC}	W_S	W_T
Fig. 1		No	√	√	√	√
		Yes		√	√	
		Biased	√	√	√	
Fig. 3	V_1	No	√	√	√	
Fig. 4	V_1	No	√	√	√	
	V_1	Yes		√	√	
	V_2	No			√	√
	V_2	Yes			√	
Fig. 5	V_2	No			√	

W_{LP} and W_{LC} indicate the pulse-front and capacitance components of W_L respectively (see Section 3.2.1).
Under extreme conditions there may be some W_{LP} and W_T components even with the reactor.

(2.3.2) Gas Pressure and Density.

In some tests the energy loss was measured for different values of reservoir heater voltage; the results were plotted in terms of gas pressure, with the use of the reservoir voltage/pressure calibration. Since the latter was determined with no arc current in the valve it may not be strictly applicable under load conditions. Of probably greater significance, however, is the consideration that the arc characteristics and energy dissipation vary with the gas density and that, even when the pressure is constant, the density varies with the temperature in the arc path. Two consequences are, first that the measured losses vary with increasing total loading, e.g. with an increase in pulse repetition frequency, in the same sense as with decreasing reservoir voltage; and secondly that the separation of losses is more difficult because the value of one component may be affected by the presence or absence of another component. For example, the leading-edge loss might be affected by the amount of inverse voltage because of the effect of the latter on trailing-edge loss and hence on valve temperature and gas density. An attempt is made to assess the effect of these considerations on the validity of the conclusions.

(2.4) Capacitance Between Anode and Cathode

At the start of conduction, the current in the valve includes a component due to the discharge of the capacitance between the anode and the cathode. This capacitance includes the inter-electrode capacitance of the valve and also the stray capacitance to earth*; in most of the experiments it was higher than in

* In the oscillogram of the pulse current the discharge of the stray capacitance is sometimes observable as a transient at the start of conduction, with a rate of rise which may be much in excess of that of the main pulse current.

normal practice. For example, in the circuit of Fig. 4 the anode-to-cathode capacitance of V_1, as measured, was 190 pF; this was made up of the inter-electrode capacitance of the valve itself, which was about 30 pF, plus that of V_2, together with capacitance contributed by the circuit including components inserted for measurement purposes. In the chopping circuit of Fig. 3(*a*) the capacitance includes that introduced by the resistor; however, the inductance of the latter delays the discharge current of the capacitance connected at the end remote from the anode, so that this has little effect on the leading-edge conditions. The presence of the saturable reactor in series with the anode has a similar effect. In the tests described below, the effective anode-to-cathode capacitance, i.e. the capacitance whose discharge current modifies the leading-edge conditions, is taken as that of the valve itself plus one-half of the capacitance to earth of the reactor, the discharge current of the other half being assumed delayed by the reactor itself. While the saturable reactor delays the rise of pulse current it has no effect on the current due to the discharge of the anode-to-cathode capacitance.

(2.5) Inverse Voltage

Arrangements were made to vary the inverse voltage appearing across the valve at the end of conduction in order to study its effect on trailing-edge loss. An inverse voltage of about 1 000 V was introduced for normal operation, by adjustment of the load to give a slight mismatch, to allow time for deionization and prevent shoot-through. Fig. 9(*a*) shows a typical oscillo-

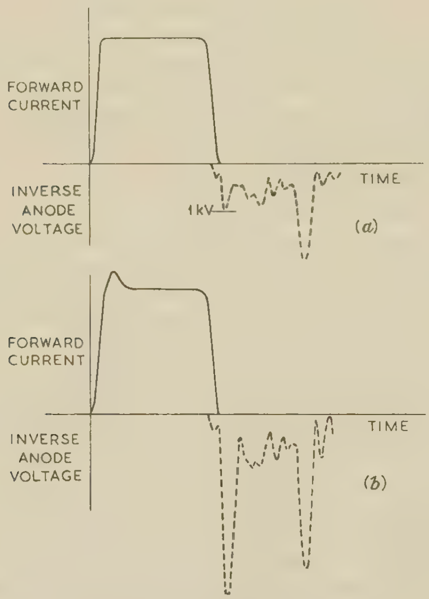


Fig. 9.—Pulse-current and inverse-voltage waveforms.
(*a*) With mismatched load.
(*b*) With added load capacitance.

gram of the voltage as measured from the anode terminal to earth. The inverse voltage appearing before the end of the current pulse is due to the rapid change of current and is developed across the inductance of the valve leads, which were unusually long because of the requirements of the experiment, as explained later; it does not appear across the anode-cathode space, the voltage here being at the arc-drop value until conduction stops.

For some experiments an inverse voltage of known magnitude and duration was applied to the valve under test at the end of conduction. In one method, a variable capacitor was connected across the load; this gave rise to an inverse voltage spik of width about 0·2 μs (at half amplitude) and of peak valu variable from 3 to 12·5 kV as shown in Fig. 9(*b*). It also cause a small current peak superimposed on the front of the curren pulse, but this occurred too late to affect leading-edge dissipation. The inverse voltage was measured by means of a voltag divider connected across the valve in series with a V1972 valv connected as an inverse diode.

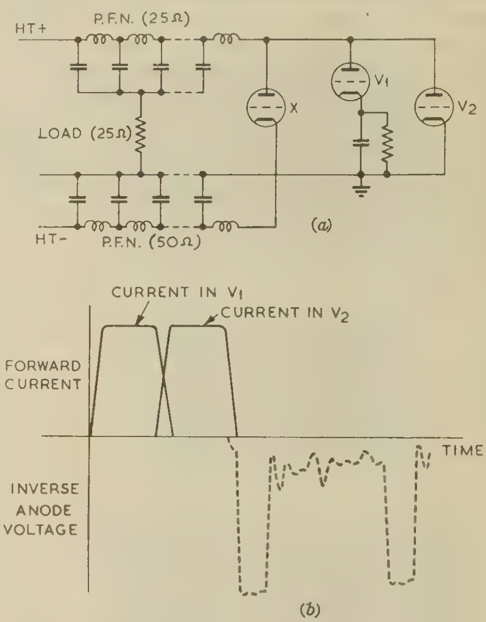


Fig. 10.—Application of an inverse voltage pulse.
(*a*) Circuit diagram.
(*b*) Current pulses in V_1, V_2 and inverse voltage across V_2.

In another method, illustrated in Fig. 10, the inverse voltag was applied by the use of a second modulator circuit. A auxiliary pulse-forming network was charged to any desire voltage (up to -18 kV) and discharged through the load b means of an auxiliary thyratron, X, to give current pulses c 1 μs duration and of variable amplitude and hence to appl corresponding inverse-voltage pulses across the second thyratro in a pulse-chopping circuit. The inverse voltage pulse wa timed to start at the end of the current pulse in the valve unde test; pulses recurred at intervals, but the second and later one could be ignored because they occurred after the residua ionization in the test valve had substantially disappeared. In th circuit used it was necessary to design the main network and th load so that their combined impedance matched that of th auxiliary network. Consequently the voltage applied in th main circuit, and as forward voltage on the valves, was mac less than in tests without the auxiliary circuit in order that th pulse current should remain the same. This did not, howeve affect the energy losses in the valve under test.

(2.6) Calorimetric Measurement

The energy dissipated in the valve was measured by means of calorimeter, the valve being completely immersed in oil and th energy loss estimated from measurements of the oil temperatur The oil container consisted of a steel cylinder, 9·5 in in diamet and 16 in long with a spiral copper tube brazed to the outsic for temperature control by water flow. The oil was continuous circulated by means of motor-driven paddles. The cylinde the paddles and the motor introduced stray capacitance betwe the anode and the cathode; this increased the energy dissipatic

in the valve, as described later. The supply leads to the thyratron were much longer than normal; a length of 6 in of each lead was immersed in the oil so that heat conducted through the leads from the thyratron should be communicated to the oil and so be measured. This introduced added inductance which was responsible for the voltage transient as the current fell to zero, as noted with reference to Fig. 9.

Both the rate of flow and the temperature of the cooling-water were closely controlled at values which restricted the change in oil temperature to a few degrees Celsius; this ensured constant conditions for heat dissipation from the valve and also avoided the rise in reservoir temperature, and hence in gas pressure, which would result from a greater increase in oil temperature. When the valve is on load the cathode is heated by the arc, the heater temperature rises and, with a constant voltage, the heater input power falls. The energy loss due to the arc is therefore equal to the value deduced from the calibration curve plus the amount of the decrease in energy supplied via the cathode heater.

(3) EXPERIMENTAL RESULTS

(3.1) Steady-State Dissipation

(3.1.1) Variation with Pulse Length.

Experimental values of W_S are given in Figs. 11 and 12. In the former case, readings of $W_L + W_S$ were made by the method

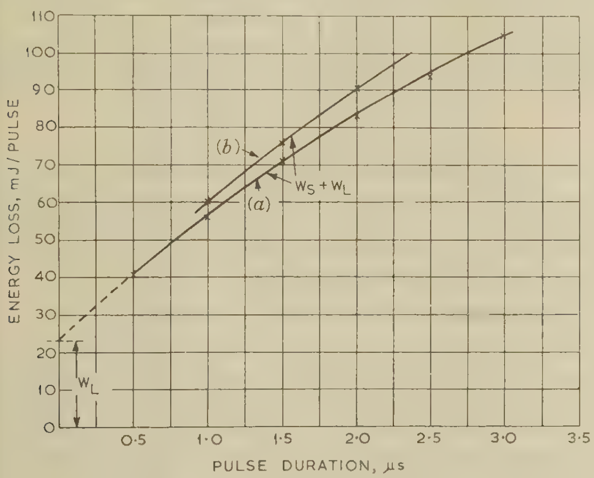


Fig. 11.—Leading-edge and steady-state loss in hydrogen as function of pulse duration.

Test conditions, 25 kV, 260 A peak, 750 pulses/sec.
(*a*) At beginning of experimental series.
(*b*) After 200 h.

of Fig. 3 ($R = 20\,\Omega$, valve V_1) for different pulse lengths. The vertical intercept on extrapolation to zero pulse length gives the value of leading-edge loss W_L (which is assumed not to vary); the value of W_S, deduced by subtracting the intercept, is seen to increase with pulse length but less than linearly.

(3.1.2) Variation with Gas Pressure and Frequency.

The curves of Fig. 12 show the steady-state loss as measured by the double-chopping method; W_S is seen to fall with decreasing pressure and with increasing pulse-repetition frequency. The shapes of the current pulses in the valves on which the tests for Figs. 11 and 12 were made were not identical but the results are substantially the same, as can be verified for the hydrogen-filled valve by comparing points on Fig. 11(*b*) with corresponding points on Fig. 12. The two curves of Fig. 11 show the results of tests made, one at the beginning and one after about 200 h

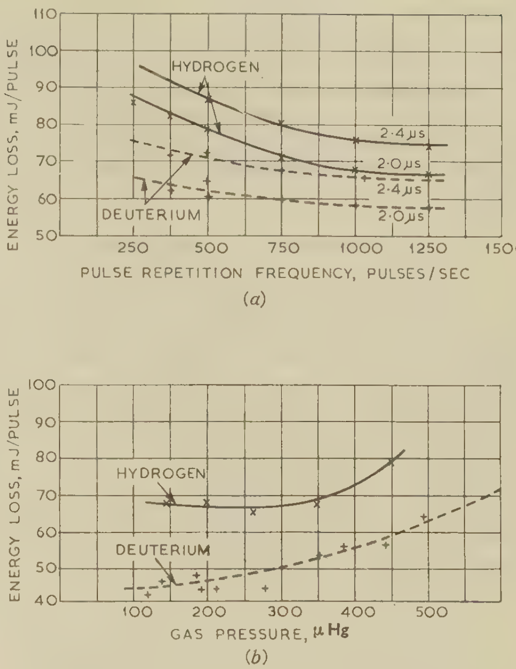


Fig. 12.—Steady-state energy loss in hydrogen and in deuterium.

(*a*) Variation of loss with pulse repetition frequency. Test conditions 260 A peak, gas pressure 450 μ Hg, pulse duration as indicated.
(*b*) Variation of loss with gas pressure.
Test conditions, 260 A peak, 500 pulses/sec, pulse duration 2 μs.

of experimental life; the energy loss increased with life, indicating an increase in arc drop. All the results quoted in the paper were from observations subsequent to the tests of Fig. 11(*b*).

(3.1.3) Variation with Peak Current.

An experiment with a saturable reactor in series with V_2 (Fig. 7) indicated the manner in which W_S increases with peak current. At 390 A, 2 μs, 500 pulses/sec the total loss, which in this case equalled the steady-stage loss, was 107 mJ/pulse. Comparing this with the loss at 260 A from Fig. 12, we observe that the increase with peak current is somewhat less than linear.

(3.1.4) Variation in Relation to Arc Drop.

These results, as well as the non-linearity of Fig. 11, are consistent with the general observation that the arc drop diminishes as the gas density is reduced; the fall in density may result either from a reduction in pressure or from a rise in temperature due to an increase in total dissipation as the peak current, the pulse length or the repetition frequency is increased.

(3.1.5) Variation with Mean Current.

Further values of W_S are obtainable by extrapolation from curves of $W_S + W_T$ described later; these, together with the results above, lead to the conclusion (discussed later) that the steady-state dissipation in watts is proportional to the mean value of the current through the valve.

(3.1.6) Variation with Gas Filling.

Fig. 12 shows that the steady-state loss in deuterium is about 20–30% lower than in hydrogen within the range of the tests. This is consistent with the fact that the arc drop is lower in deuterium.[4]

(3.1.7) Validity of Calculations.

The values of W_S thus measured are used to deduce values of W_L and W_T from measured values of $W_L + W_S$ and $W_S + W_T$.

This assumes that W_S as measured alone is the same as it would be if W_L or W_T were present in the same valve (thus increasing the total loss). The agreement between Figs. 11(*b*) and 12 supports the assumption within the limits of the tests; so also does a test, described later, in which the total losses are measured and plotted as a function of frequency, the curve agreeing closely with that obtained by a summation of the loss components as measured separately.

(3.2) Leading-Edge Dissipation

(3.2.1) Two Components of Loss.

We may distinguish two components of leading-edge dissipation associated with two components of current as the valve voltage falls to its steady arc-drop value:

(*a*) Pulse-front loss, i.e. the energy dissipation due to the current from the discharge of the pulse-forming network.
(*b*) Capacitance component, due to the current from the discharge of the anode-to-cathode capacitance.

The significance of (*b*) was studied by measurements of the combined loss $W_L + W_S$ in V_1 of the pulse-chopping circuit of Fig. 4. Measurements of dissipation were made with different capacitances added by connection between the anode and cathode of the valve. Fig. 13 shows the measured loss as a function of total capacitance. By extrapolation we have, for zero added capacitance, i.e. for a total of 30 pF (the inter-electrode capacitance of the valve), a dissipation of 88 mJ/pulse; this is in close agreement with the sum of the energy stored in this capacitance ($\frac{1}{2}CV^2 = 9 \cdot 5$ mJ) and the steady-state loss of 76 mJ given by Fig. 12(*a*).

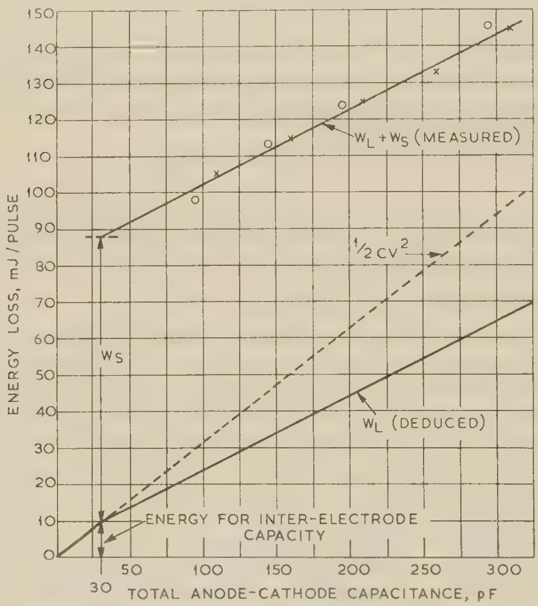


Fig. 13.—Effect of anode–cathode capacitance on leading-edge loss in hydrogen.

Test conditions, 25 kV, 260 A peak, 2·4 μs, 1 000 pulses/sec.
× Without saturable reactor.
○ With saturable reactor.

Thus it appears that at normal gas pressures almost the whole of the leading-edge dissipation is due to the discharge of the energy stored in the anode-to-cathode capacitance. This conclusion is supported by similar tests with the saturable reactor in circuit. The results, also plotted in Fig. 13, show that the reactor had negligible effect on the losses recorded; these cannot then have included any pulse-front component since this would have been eliminated by the reactor. Again, any component of leading-edge loss due to the pulse current might be expected to vary with the rate of rise of current, but tests showed that at normal gas pressure the loss rose by less than 15% for an increase in average rate of rise (between 25 and 75% of peak value) from 1 000 to 6 500 A/μs.

The variation of leading-edge loss with pulse frequency at normal gas pressure is shown in Fig. 14. It was computed by measuring the total leading-edge and steady-state losses in V_1 in the circuit of Fig. 4, and subtracting the appropriate values of the latter as given in Fig. 12(*a*). The leading-edge loss is seen to increase with frequency; this may be due to the reduction in gas density as the total loading is increased.

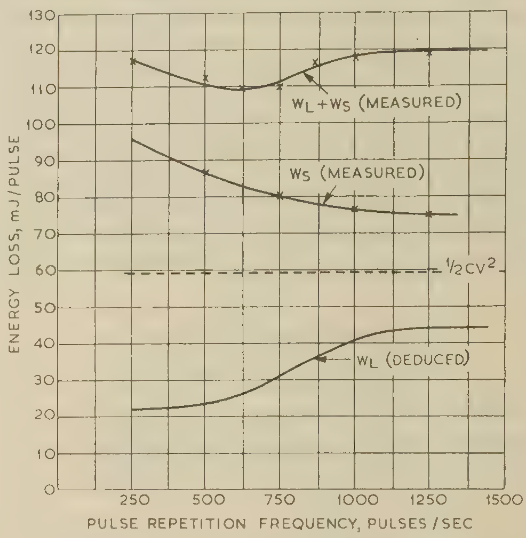


Fig. 14.—Leading-edge and steady-state loss in hydrogen as function of pulse-repetition frequency.

Test conditions, 25 kV, 260 A peak, 2·4 μs.

(3.2.2) Anode Heating Due to Capacitance Discharge.

In the tests of Fig. 13 the anode was seen to be red hot when the anode-to-cathode capacitance was 310 pF. In these conditions the power associated with the component W_L was 67 W. The anode also became red in the tests of Fig. 14 at a frequency of 1 500 pulses/sec when the power loss due to W_L was 68 W. Here the anode-to-cathode capacitance was 190 pF; in a further test in which this was reduced to 45 pF, anode heating was not visible at frequencies up to 2 000 pulses/sec.

(3.2.3) Discharge of the Anode-to-Cathode Capacitance.

The leading-edge loss in the tests recorded in Fig. 14 is mainly that due to the discharge of the anode-to-cathode capacitance, which was 190 pF (as measured with the pulse-forming network, including its end inductance, disconnected from the anode circuit). However, the highest value of W_L recorded is only 80% of the energy, $\frac{1}{2}CV^2$, stored in the added capacitance and is in most cases much less than this. The reason may be, either that the capacitance is not completely discharged, the degree of discharge varying with the gas density and being greater the slower the fall of anode-cathode voltage, or that a variable amount of the energy is dissipated in leads and other circuit components and does not appear in the thyratron. The dissipation recorded is never less than that corresponding to the inter-electrode capacitance of the valve (30 pF in this case), suggesting that this component at least is dissipated within the valve.

Tests were made in a separate circuit in which a 1 000 pF

apacitor was connected directly between the anode and cathode f the same thyratron as was used for the other tests described. 'he capacitor was charged to 15 kV through a resistance chain nd discharged at various repetition frequencies, the dissipation 1 the valve being measured by means of the calorimeter. Table 2 ndicates some results obtained.

Table 2

Repetition frequency	Hydrogen pressure	Measured dissipation
pulses/sec	μ Hg	mJ/pulse
250	450	88
500	450	94
750	450	97
500	145	116

The energy, $\frac{1}{2}CV^2$, stored in the capacitance at each pulse is 12·5 mJ. The dissipation as recorded is seen to increase for ither an increase in repetition frequency or a decrease in gas ressure, as in the modulator tests, but the value is in all cases a igher proportion of the theoretical value than in the modulator ircuit.

3.2.4) Variation of Leading-Edge Loss with Pressure.

The above observations show that at normal gas pressures 450 μ Hg and above) the pulse component of leading-edge loss s negligible; this indicates a rapid fall of the anode-cathode 'oltage to the steady arc-drop value. The case is different, iowever, at low gas pressures. Fig. 15 shows how the leading-

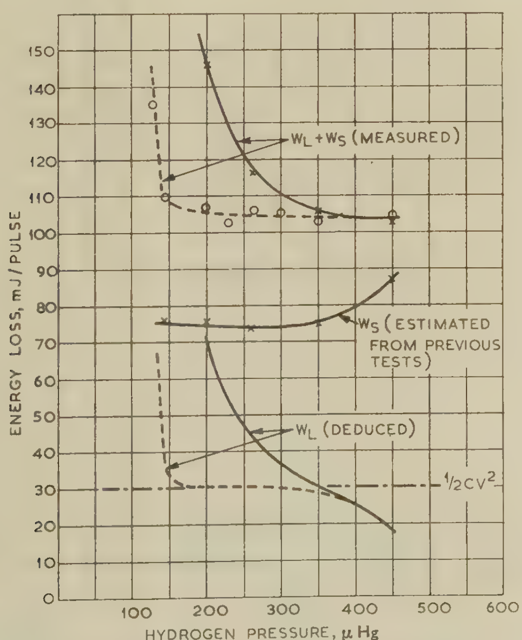


Fig. 15.—Effect of hydrogen pressure on leading-edge loss.

Test conditions, 25 kV, 260 A peak, 2·4 μs, 500 pulses/sec.
—— Without saturable reactor.
– – – With saturable reactor.

edge loss in hydrogen increases as pressure falls, becoming much iigher than the energy stored in the anode-to-cathode capaci-ance. As the pressure falls, the pulse-front component of loss ecomes significant at about 350 μ Hg, increasing rapidly there-fter; but it can be very largely eliminated by the use of a saturable reactor, the allowable working pressure range thus being extended down to about 150 μ Hg.

Measurements made in the deuterium-filled valve are recorded in Fig. 16. They confirm the above observations as to the

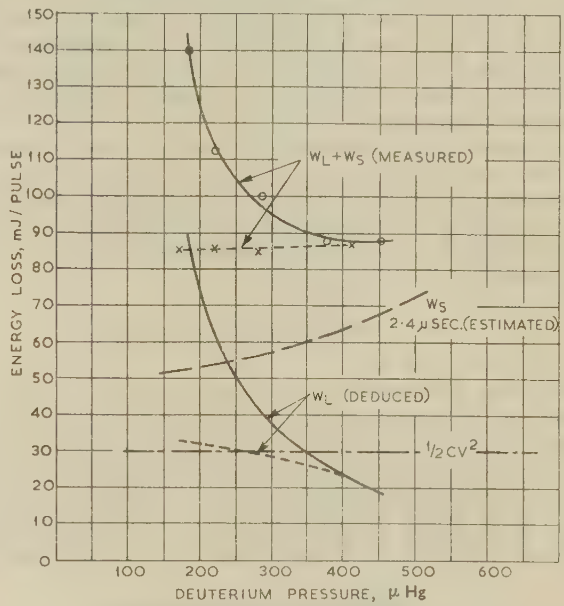


Fig. 16.—Effect of deuterium pressure on leading-edge loss.

Test conditions, 25 kV, 260 A peak, 2·4 μs, 500 pulses/sec.
—— Without saturable reactor.
– – – With saturable reactor.

capacitance and pulse-front components; at low gas pressures the latter is much the same in both gases. This is as may be expected from the observed fact that, for the same gas pressure, the rate of fall of the voltage between anode and cathode is practically the same in deuterium as in hydrogen.[5]

(3.2.5) Variation of Leading-Edge Loss with Repetition Frequency.

From the pulse-chopping tests it appeared that at normal pressures the pulse component of leading-edge loss was negligible at repetition frequencies up to 1 500 pulses/sec. With the saturable reactor it was possible to make tests at higher frequencies. The results of two tests at normal gas pressures are shown in Fig. 17. In one case the reactor eliminated the trailing-edge loss

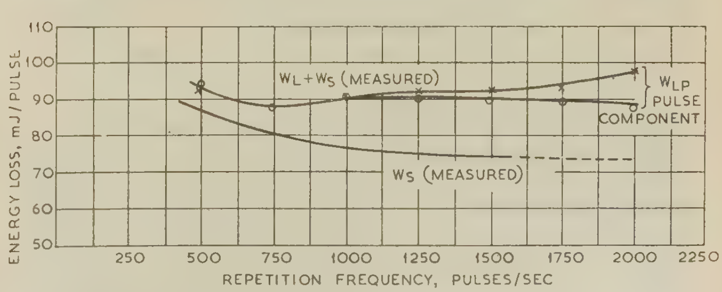


Fig. 17.—Effect of repetition frequency on leading-edge loss in hydrogen.

Test conditions, 25 kV, 260 A peak, 2·4 μs, 450 μ Hg.

W_T and the pulse component of W_L; in the other it was biased so as to eliminate W_T only. The loss W_S and the capacitance component of W_L were unaffected by the presence of the reactor; these losses were recorded in one test, and these plus the pulse

component of W_L in the other. The difference between the two curves corresponds to the latter component; this component, which is negligible (at this gas pressure) at frequencies up to 1 500 pulses/sec, is seen to be significant at higher frequencies, increasing with frequency. The effect is presumably due to the reduced gas density under the higher total loading (comparable with the effect at reduced gas pressure). The contribution of this component to anode heating, even at normal gas pressure, may thus be a factor in limiting the permissible rating at these and higher frequencies.

The value of the capacitance component of W_L, computed from these curves by subtracting values of W_S (from Fig. 12), is approximately equal to the energy, $\frac{1}{2}CV^2$, corresponding to the anode-to-cathode capacitance, which is estimated at 45 pF (30 pF for the thyratron plus 15 pF representing one-half of the stray capacitance to earth of the reactor).

(3.3) Trailing-Edge Dissipation

(3.3.1) Inverse Current.

The energy involved in the passage of inverse current is a function of the amplitude and duration of the inverse voltage and also of the concentration of residual ionization and of its time of persistence. Measurements were made with inverse voltages of variable amplitude, applied as described with reference to Figs. 9 and 10. Oscillograms of the inverse current, Fig. 18, show that for an inverse-voltage pulse of given amplitude

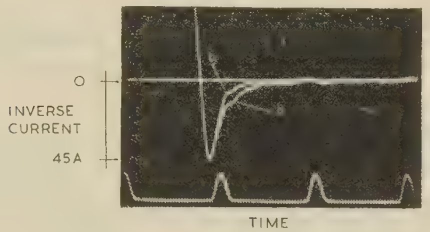


Fig. 18.—Superimposed inverse current oscillograms for different hydrogen pressures.

Test conditions, short-circuit load.
Calibration marks at 1 μs intervals.
(*a*) Hydrogen pressure, 200 μ Hg.
(*b*) Hydrogen pressure, 450 μ Hg.

the peak current does not vary with pressure (within the usual range of operation) but that the time of persistence of current increases with pressure, as might be expected from the reduction in the rate of deionization. The measurements were made in the circuit of Fig. 1 but with the load short-circuited, the charging voltage being adjusted to give a forward-current pulse of 300 A with 2 μs duration.

(3.3.2) Variation with Gas Pressure and Repetition Frequency.

Tests were made in the circuit of Fig. 10, with an applied inverse voltage of 6·7 kV peak and 1 μs duration, to measure the trailing-edge loss, W_T as a function of gas pressure and of repetition frequency. The total loss measured in the valve under test was the trailing-edge loss plus steady-state loss; the latter being known from Fig. 12, the former could be calculated by subtraction. The results are shown in Fig. 19. In each case a decrease in dissipation accompanies a reduction in gas density because of the more rapid deionization.

(3.3.3) Variation with Amplitude and Duration of Inverse Voltage.

Fig. 20 records similar tests at constant frequency and pressure, showing the variation of $W_S + W_T$ with the amplitude of the inverse voltage. Extrapolation to the axis for zero inverse

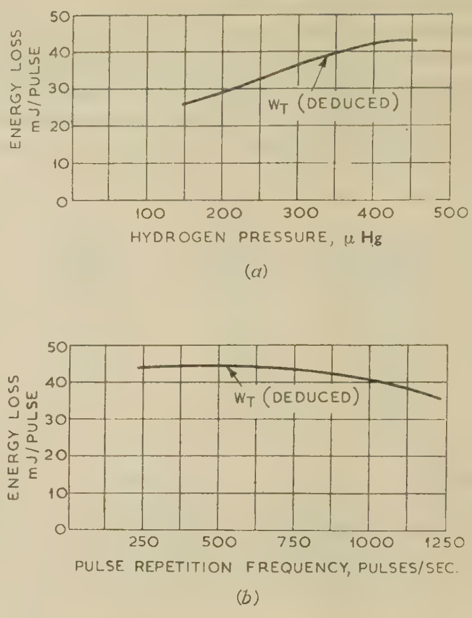


Fig. 19.—Trailing-edge energy loss in hydrogen.

(*a*) Variation of loss with hydrogen pressure.
Test conditions, 260 A peak, 2 μs, 500 pulses/sec. Rate of fall of current 1 200 A/μs.
Inverse voltage 6·7 kV, duration 1 μs.
(*b*) Variation of loss with pulse repetition frequency.
Test conditions, 260 A peak, 2 μs, 450 μ Hg. Rate of fall of current, 1 200 A/μs.
Inverse voltage, 6·7 kV, duration 1 μs

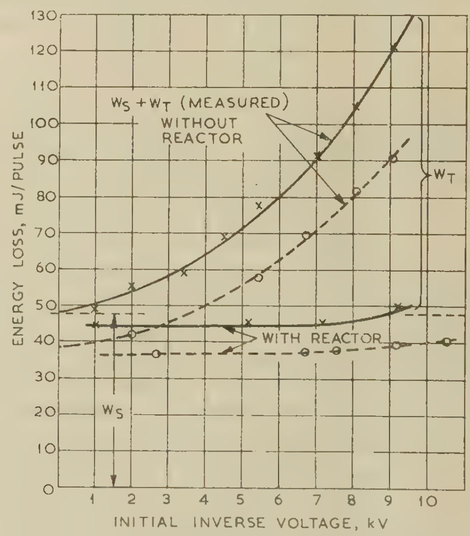


Fig. 20.—Steady-state and trailing-edge loss in hydrogen and in deuterium as function of inverse voltage.

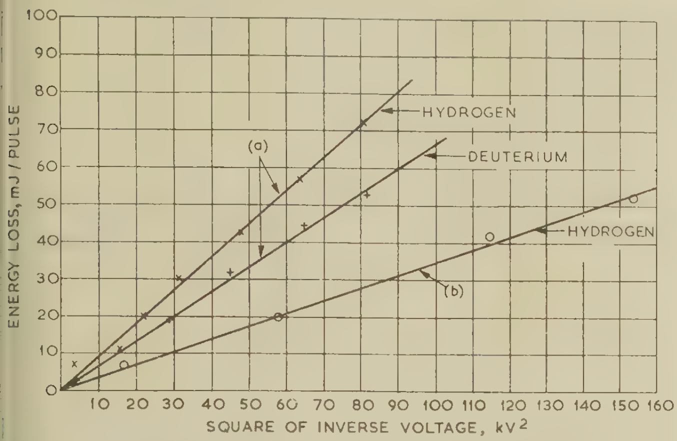
Test conditions, 260 A peak, 1·15 μs, 500 pulses/sec, 450 μ Hg. Rate of fall of current 1 200 A/μs.

—×— Hydrogen.
—○— Deuterium.

voltage gives the steady-state loss; for the hydrogen-filled valve this gives a value of 48 mJ/pulse, which agrees closely with the value which can be calculated from Figs. 11(*b*) and 12. With a saturable reactor in circuit the trailing-edge loss is seen to be practically eliminated, beginning to be significant only at high values of inverse voltage. The observations with deuterium are similar to those with hydrogen, except that the trailing-edge loss is about 25% lower in deuterium.

Very similar results were obtained in tests in hydrogen with a variable inverse-voltage spike of only 0·2 μs duration [Fig. 9(*b*)]

n each case the trailing-edge loss may be calculated by subtracting from the total the intercept on the vertical axis. The values f W_T computed in this manner are replotted in Fig. 21; the



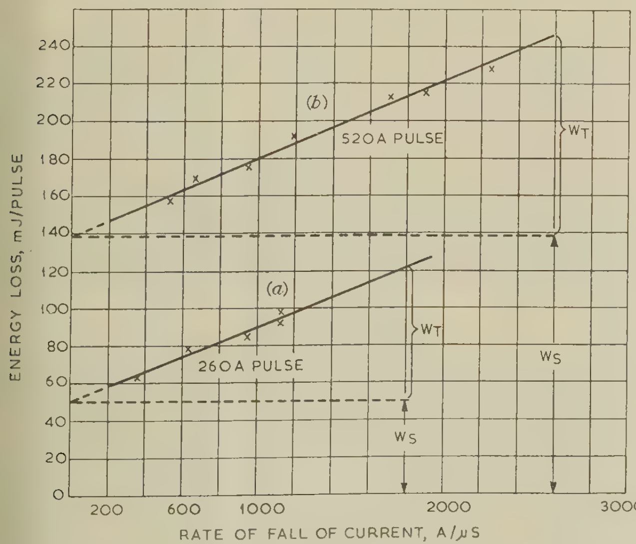
ig. 21.—Trailing-edge loss in hydrogen and in deuterium as function of square of inverse voltage

Test conditions, 260 A peak, 1·15 µs, 500 pulses/sec, 450 µ Hg. Rate of fall of current 1 200 A/µs.
(a) Inverse voltage duration, 1 µs.
(b) Inverse voltage duration, 0·2 µs.

urves show that the trailing-edge dissipation varies directly with he square of the amplitude of the inverse voltage. It also is een to increase with the duration of the inverse voltage, but the ncrease would not extend beyond about 1 μs, after which eionization appears to be effectively complete, as seen from Fig. 18.

3.3.4) Variation with Current Peak and Rate of Fall.

The trailing-edge loss depends on the ionization remaining when forward conduction stops. Loss measurements were made, similar to those recorded in Fig. 20, but with a fixed value of inverse voltage and with a variable rate of fall of current at the end of the pulse. The observations are plotted in Fig. 22 for two different peak currents, 260 and 520 A; they show that



ig. 22.—Trailing-edge loss in hydrogen as function of rate of fall of current.

Test conditions:
(a) 260 A peak, 1·15 µs, 500 pulses/sec, 450 µ Hg. Inverse voltage, 6·7 kV duration 1 µs.
(b) 520 A peak, 2 µs, 500 pulses/sec, 450 µ Hg. Inverse voltage 6·7 kV, duration 1 µs.

the trailing-edge loss is directly proportional to the rate of fall of current and, within the range of the tests, independent of the peak current.

(3.4) Consistency of Methods of Measurement

The methods used for the measurement of loss are justified by the similarity in the results obtained in different experiments, including those in which the value of a component of loss is estimated by extrapolation. A further check was made by measuring the total loss in a single valve under specified conditions and without current chopping. Fig. 23 shows how the

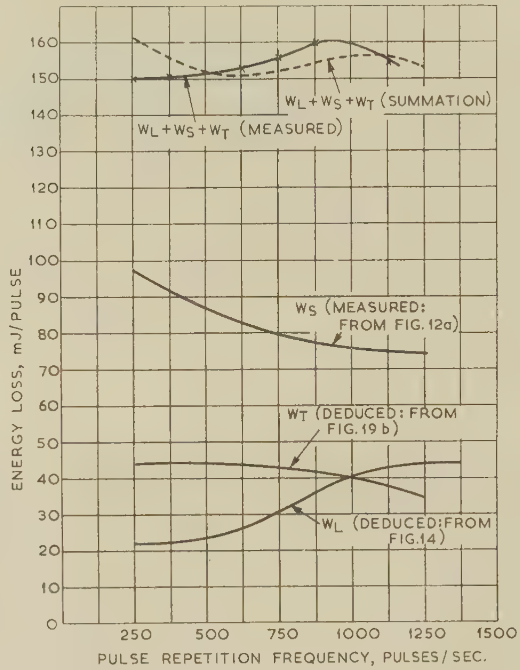


Fig. 23.—Total energy loss in hydrogen as function of pulse frequency.

Test conditions, 25 kV, 260 A peak, 2·4 µs, 450 µ Hg. Rate of fall of current 1 200 A/µs. Inverse voltage, 6·7 kV, duration 1 µs.

total loss varied with repetition frequency. On the same graph are plotted curves for the separate loss-components as measured in previous experiments and also a curve showing the summation of these three. The latter is seen to be in good agreement with the curve of measured total dissipation.

(3.5) Mode of Energy Dissipation

In the discharge, energy is dissipated by the transfer of the energy of accelerated particles to the electrodes, by the release of the energies of recombination and of reassociation at electrodes and at walls, and by conduction by the heated gas molecules. Some tentative conclusions as to the mode of energy dissipation may be drawn from the changes in temperature of the cathode heater and from the visible heating of the grid and the anode. When the heater input (at constant voltage) falls because of heating of the cathode by the arc, the effect is apparently due to the steady-state loss W_S: it is the same whether or not W_S is accompanied by a component of either W_L or W_T. In valves subject to W_S only (as in V_2 of Fig. 5), there was no visible heating of either the anode or the grid. With W_L and W_S together, the anode may be at bright-red heat with no visible heating of the grid. On the other hand, with W_T and W_S together the first sign of heating is at the grid.

These observations suggest that practically all of W_S is dissipated at the cathode, practically all of W_L at the anode, and at

*

least a substantial part of W_T at the grid. At normal pressure the component W_L is mainly due to the discharge of the anode-to-cathode capacitance. At lower pressures there is a pulse-front component; anode overheating occurs at lower power than at normal pressures because the heat conductivity of the gas falls with its density and the effect is cumulative as the anode temperature rises, giving a runaway effect with the anode suddenly reaching very high temperatures. In the tests of Figs. 14 and 23 the anode was red hot at frequencies of 1 500 and 1 125 pulses/sec respectively; in the former case the dissipation rate corresponding to W_L was 68 W and in the latter it was 49 W for W_L and 42 W for W_T. This is consistent with the approximation that all of W_L is dissipated at the anode and that one-half of W_T is dissipated at the anode and one-half at the grid.

The tests were repeated to find the frequency in each case at which the anode was just visibly red, the tests being performed in quick succession to enable a comparison to be made from appearance. The frequencies were 1 300 pulses/sec in the first case, giving a rate of 58 W for W_L, and 1 000 pulses/sec in the second case, giving 41 W for W_L and 39 W for W_T. This also is in agreement with the rough approximation given above.

The part played by the capacitance component of W_L is confirmed by the observation that the anode was red hot for the last point plotted on Fig. 13 for a capacitance of 310 pF; the deduced value of W_L in this case was 67 mJ/pulse, giving a dissipation rate of 67 W.

(3.6) Conclusions

Except at low gas pressures, the leading-edge loss is predominantly that due to the discharge of the anode-to-cathode capacitance, which should therefore be reduced to a minimum; the loss varies with the square of the anode voltage and equals a fraction (varying with operating conditions) of the charge energy of the capacitance. At low pressures the leading-edge loss includes a pulse-front component which increases rapidly with decreasing pressure. This component may also be significant when the gas density is reduced by valve heating under high total loading. The use of a saturable reactor does not affect the capacitance component of loss but it may largely eliminate the pulse-front component, enabling the thyratron to operate satisfactorily at pressures much lower than could otherwise be allowed.

The steady-state loss increases with the current amplitude and with pulse length, but less than linearly, because the arc drop falls as the total dissipation is increased. Within the limitations of the experiments, the loss is proportional to the arithmetic mean value of the pulse current. This is illustrated by Fig. 24, which plots the steady-state loss in watts as a function of mean current for peak values of 260, 390 and 520 A, as measured in various experiments described above.

The trailing-edge loss increases directly with the rate of fall of the pulse current, with the square of the inverse-voltage amplitude and with the duration of this voltage (up to the effective completion of deionization, which, in these tests, generally took less than 1 μs). In practical circuits it may constitute an important fraction of the total arc loss: it can, however, be eliminated to a considerable extent by the use of a saturable reactor.

The steady-state loss is dissipated mainly at the cathode, the leading-edge loss mainly at the anode, and the trailing-edge loss about half at the anode and half at the grid. The anode of the BT101 valve under test became red hot when energy was released at it by the two latter components at a rate of about 60 W.

Each component of loss varies similarly with operating conditions in both hydrogen and deuterium. The two components of leading-edge loss are much the same in each gas but

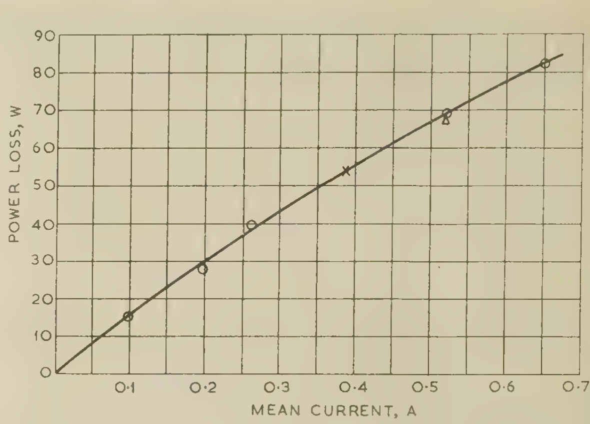


Fig. 24.—Steady-state power loss in hydrogen as function of mean current.

○ 260 A. × 390 A. △ 520 A.

the steady-state and trailing-edge losses are 20–30% lower in deuterium.

In the rating specification of thyratrons for modulator duty, the quotation of a maximum permissible value of mean current is consistent with the observation that the steady-state loss, generally the largest of the three components, is proportional to this value. The operation factor, the maximum allowed value of the product of forward voltage, peak current and repetition frequency, limits a component of loss which is generally negligible at normal pressures, namely the pulse component at the leading edge; it has little relation to the capacitance component, which is often more important. The ratings should make reference to maximum allowable values of anode-to-cathode capacitance.

A typical specification for inverse voltage is that 'except for a spike of 0·05 μs duration the peak inverse voltage shall not exceed 5·0 kV during the first 25 μs after the pulse'. In practice the inverse voltage commonly exceeds this allowance and introduces an important component of loss. If trailing-edge losses are incurred they should be compensated for by appropriate limitation elsewhere, for example in the mean current.

(4) ACKNOWLEDGMENTS

The work described was carried out in the Research Laboratory, A.E.I. (Rugby) Ltd., and the authors thank their Director for permission to publish this account. They acknowledge the help, both in technical discussion and in active collaboration, of their colleagues D. H. Grindell and S. W. Longstaff and of a former colleague, N. R. McCormick, and also of members of the Royal Radar Establishment, and in particular N. S. Nicholls: thanks are due to the latter for the circuit of Fig. 10 and to A. Langley Morris for the design and construction of the saturable reactor.

(5) REFERENCES

(1) EICHENAUER, C.: Fourth Symposium on Hydrogen Thyratrons, Fort Monmouth, N.J., U.S.A., 17th November 1955.

CREEDON, J. E., and SCHNEIDER, S.: Fifth Symposium on Hydrogen Thyratrons, Fort Monmouth, N.J., U.S.A., 20th May 1958.

(2) SCHMIDT, A.: *Transactions of the American I.E.E.*, 1946, 65, p. 654.

(3) GRAFF-BAKER, W. S., and STOREY, D. J.: British Patent Specification 738002.

(4) COOK, K. G., and ISAACS, G. G.: *British Journal of Applied Physics*, 1958, 9, p. 497.

(5) ARMSTRONG, R. J., and NICHOLLS, N. S.: *ibid.*, p. 498.

21.313.323 : 621—52

The Institution of Electrical Engineers
Paper No. 3594 M
July 1961
©

A SIMPLE ELECTRODYNAMIC MULTIPLIER USING TORQUE BALANCE

By J. G. HENDERSON, B.Sc., Associate Member.

(*The paper was first received* 9*th December*, 1960, *and in revised form* 13*th March*, 1961.)

SUMMARY

An electrodynamic multiplying system is described which uses simple limited-range torque motors in a torque-balance feedback system. Certain inherent errors exist and the conditions for keeping these within tolerable limits are indicated.

(1) INTRODUCTION

In the field of analogue computation and industrial automatic control there is a need for a simple, robust, accurate and cheap multiplying device. With the advent of adaptive types of control schemes based on optimization criteria such as the minimization of squared error, there is likely to be a growing demand for such multipliers.

The multiplier to be described* is essentially an electrodynamic torque-balance feedback system after the style of a pair of opposed electrodynamic-wattmeter elements, but instead of air-cored coils the torque-producing elements are torque motors of the Microsyn pattern, as in Fig. 1. In these units all the operating coils are located on the stator and there are no mechanical connections to the rotor. In consequence the device is one with low shaft friction, high torque/inertia ratio and low power consumption (about 10 mW/coil), which makes it ideal for transistorized operation.

Tests on a prototype have shown a multiplication accuracy of $\pm 1{\cdot}5\%$ with excellent zero stability. The cut-off frequency of the feedback loop is 150 c/s and hence the device can be used as an accurate analogue multiplier over a frequency band of about 0–30 c/s. This dynamic performance is intermediate between that of multipliers using servo-driven voltage dividers whose wear is a limitation for continuous operation, and the much greater bandwidth of all electronic multipliers which are, however, more elaborate. The small size, simplicity, robustness and stability of this multiplier may make it an attractive proposition for industrial instrumentation on continuous processes.

(2) PRINCIPLE OF OPERATION

The electrodynamic element of the multiplier consists of two torque motors and an a.c. position pick-off enclosed in a common tube with their three salient-pole rotors mounted on a common shaft which is held in instrument-type jewelled bearings, as shown in Fig. 1(*a*).

The two currents to be multiplied, i_1 and i_2, are fed into torque motor 1, Fig. 2, which produces a torque, T_1, proportional to $i_1 i_2$ tending to turn the rotor system. This motion is detected by the centre unit, which is an a.c. position pick-off working with a carrier frequency of 5 kc/s, and the output voltage, proportional to angular deviation θ, is fed into an a.c. amplifier. Subsequently this signal is passed through a phase-sensitive rectifier and a d.c. amplifier and the resultant output current, i_3, is then fed into one winding of torque motor 2, the other winding of

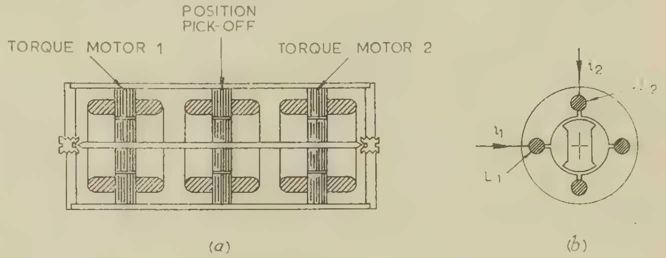


Fig. 1.—General arrangement of electrodynamic multiplying unit.
(*a*) Longitudinal arrangement of torque motors and position pick-off.
(*b*) Cross-section through torque motor.

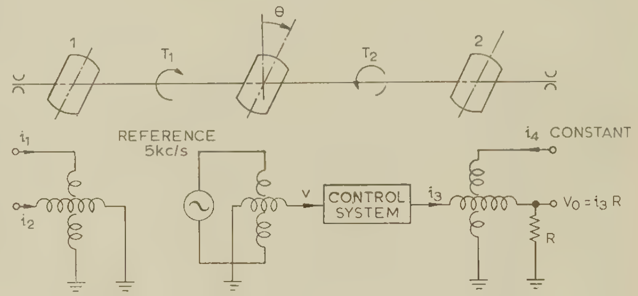


Fig. 2.—Schematic of torque-balance multiplier.

which is fed with a constant polarizing current, i_4. The resultant torque T_2, proportional to $i_3 i_4$, is arranged to oppose T_1 and hence by this negative feedback, $T_1 \simeq T_2$. Thus, the output current is $i_3 \simeq i_1 i_2 / i_4$ and therefore the device can be used as a multiplier.

(2.1) Principle of the Torque Motor

The ideal construction of a limited-range torque motor is shown in Fig. 1(*b*), where it can be seen that the two operating coils, L_1 and L_2, are at right angles and that the rotor is of simple salient-pole construction. The laminations are of Radiometal or Mumetal, in order to minimize errors due to hysteresis and remanence. In general, the torque developed by such a motor is

$$T = \frac{1}{2} i_1^2 \frac{dL_1}{d\theta} + \frac{1}{2} i_2^2 \frac{dL_2}{d\theta} + i_1 i_2 \frac{dM_{12}}{d\theta} \quad . \quad . \quad . \quad (1)$$

$$= T_a + T_b + T_c$$

where $dL_1/d\theta$, $dL_2/d\theta$ and $dM_{12}/d\theta$ are the angular rates of change of the winding self- and mutual inductances.

The multiplier depends for its operation upon the third term, T_c, and hence for accurate multiplication T_a and T_b must be small enough in comparison with T_c for the required accuracy to be achieved. It should be noted that in the case of a system using air-cored coils, like those of an electrodynamic wattmeter, the T_a and T_b terms are zero and the torque of such units is

* Provisional Patent No. 4977 (1960).

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Henderson is in the Department of Electrical Engineering, University of Birmingham.

simply $i_1 i_2 dM_{12}/d\theta$. Torque-balance multipliers of this kind* are capable of the highest accuracy, but they tend to be large and the power consumption of the operating coils is high.

The relative magnitude of the torque terms, T_a and T_b in eqn. (1), can be assessed from the variations of L_1, L_2 and M_{12} with angular movement, θ, of the rotor. The inductance L_1 is a maximum when $\theta = 0$, 180° and 360°, and a minimum at $\theta = 90°$ and 270°; thus there is a periodic variation of L_1 of the general form

$$L_1 = L_a + L_b \cos 2\theta \quad . \quad . \quad . \quad . \quad . \quad (2)$$

and if L_1 and L_2 have the same number of turns,

$$L_2 = L_a - L_b \cos 2\theta \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The mutual inductance variation will be of the form

$$M_{12} = M \sin 2\theta \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where $M \simeq L_a$. Thus

$$\frac{dL_1}{d\theta} = -2L_b \sin 2\theta \quad . \quad . \quad . \quad . \quad . \quad (5)$$

$$\frac{dL_2}{d\theta} = 2L_b \sin 2\theta \quad . \quad . \quad . \quad . \quad . \quad (6)$$

and

$$\frac{dM_{12}}{d\theta} = 2M \cos 2\theta \quad . \quad . \quad . \quad . \quad . \quad (7)$$

We may assume $L_b \simeq L_a/2$ and if $i_{1\,max} = i_{2\,max}$ we can determine the relative magnitudes of the disturbing terms:

$$\frac{T_a}{T_c} = -\frac{L_b \sin 2\theta}{2M \cos 2\theta} \rightarrow -\frac{\theta}{2} \text{ as } \theta \rightarrow 0 \quad . \quad . \quad . \quad (8)$$

and

$$\frac{T_b}{T_c} = +\frac{L_b \sin 2\theta}{2M \cos 2\theta} \rightarrow \frac{\theta}{2} \text{ as } \theta \rightarrow 0 \quad . \quad . \quad . \quad (9)$$

Thus, for the absolute values of T_a and T_b to be less than 1% of T_c, the angular movement of the rotor must be kept within $\theta = \pm 0{\cdot}02$ rad, i.e. approximately $\pm 1°$. Actually, T_a and T_b are of opposite sign and therefore slightly greater displacements may be allowed.

The torque motors used in the prototype were made from 12-slot size 11 synchro laminations which were wound with two coils across both the horizontal and vertical axes. These distributed windings give inductance variations of the form assumed. However, torque motors of the type shown in Fig. 1(*b*) with concentrated windings may be expected to give smaller values of $dL_1/d\theta$ and $dL_2/d\theta$ and an improved performance is expected with these.

The magnitude of the torque, T_c, can be specified in terms of the torque-motor dimensions:

$$T_c = i_1 i_2 \frac{dM_{12}}{d\theta} = i_2 \frac{d\psi_{21}}{d\theta} \text{ newton-meters} \quad . \quad . \quad (10)$$

* Estoppey, R. F.: 'The Inductronic Electrodynamometer for the Precise Measurement of Voltage, Current, Power and Energy', *Transactions of the American I.E.E.*, 1959, **78**, Part 1, p. 393.

In this equation,

$$\frac{d\psi_{21}}{d\theta} = \text{angular rate of change of mutual flux-linkages}$$

$$= \mu_0 \frac{i_1 N_1}{2g} 2rlN_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

where r = Mean radius of air gap, m.
g = Air-gap length, m.
l = Axial length, m.

Thus,

$$T_c = i_1 N_1 i_2 N_2 \frac{\mu_0 r l}{g} \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

Normally the coil m.m.f.s, $i_1 N_1$ and $i_2 N_2$, are equal and are chosen so that the flux density in the iron circuit is about 50% of the value at which the magnetic circuit begins to saturate.

The torque motors used in the prototype had dimensions $r = 0{\cdot}7$ cm, $l = 0{\cdot}4$ cm and $g = 0{\cdot}02$ cm. The windings each had 1000 turns and hence with maximum currents of 20 mA these gave a rated value of $T_c = 700$ dyne-cm. The moment of inertia of the rotor was $0{\cdot}74\,\text{g-cm}^2$. The winding resistances were $150\,\Omega$ each and in association with winding inductances of 70–100 mH these gave an intrinsic winding time-constant of the order of 0·6 ms. In practice, however, when using voltage sources for test purposes it is quite convenient to include a swamp resistance in order to raise the input impedance, and all the tests described were made with total circuit resistances of $1\,\text{k}\Omega$.

(2.2) Torque-Motor Dynamics and Control-System Performance

The general dynamic scheme of the multiplier is shown in Fig. 3, where it can be seen that the actuating torque, T, is the difference between the input torque, T_1, proportional to $i_1 i_2$, and the output torque, T_2, proportional to the product of the

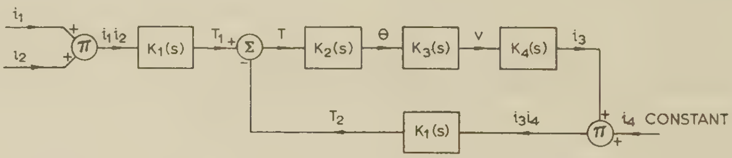


Fig. 3.—Block diagram of the multiplier.

control-system output current, i_3, and the constant polarizing current, i_4.

In this diagram K_1 is the torque-motor constant, substantially the same for both motors and equal to $dM_{12}/d\theta$. The effect of the small and variable error torques due to $dL_1/d\theta$ and $dL_2/d\theta$ are neglected in this diagram. However, the effect of the constant polarizing current, i_4, is not negligible as it gives the moving system of the basic transducer a small restoring torque in response to angular displacements θ though the agency of the term $\frac{1}{2} i_4^2 dL_4/d\theta$. Thus the transfer-function $\frac{\theta}{T}(s) = \mathrm{K}_2(s)$ is the simple quadratic-function characteristic of an oscillatory system, namely

$$\mathrm{K}_2(s) = \frac{K_2}{s^2 + 2\alpha s + \omega_0^2}$$

where the undamped natural frequency, ω_0, is dependent upon the above torsional stiffness and the moment of inertia of the system.

The shaft position pick-off is of the same pattern as the torque

motors, and when fed with a 5 kc/s voltage it gives an output voltage the magnitude of which is proportional to θ and the phase of which reverses with the sign of θ. Hence the pick-off transfer-function is a scalar, K_3.

Thus, if we designate the response of the control circuit as $K_4(s)$, the overall system response will be

$$\frac{i_3}{i_1 i_2}(s) = \frac{K_1(s)K_2(s)K_3(s)K_4(s)}{1 + K_1(s)K_2(s)K_3(s)K_4(s)} \frac{1}{i_4}$$

and, as is well known, provided that

$$K_1(j\omega)K_2(j\omega)K_3(j\omega)K_4(j\omega) \gg 1$$

for low frequencies, we shall obtain the desired condition that $i_3 \simeq i_1 i_2 / i_4$.

The control system can be designed once the dynamic response of the basic transducer $K_1(s)K_2(s)K_3(s)$ is known, and this can be found quite simply by breaking the feedback path and measuring the response of pick-off voltage to a variable-frequency test current, i_3. This characteristic is shown in Fig. 4(*a*). It can be seen that, with the addition of a simple phase-advancer centred on 125 c/s, a phase margin of about 30° can be obtained with a system attenuation of 52 dB and thus the zero-frequency gain of the system can be about 300.

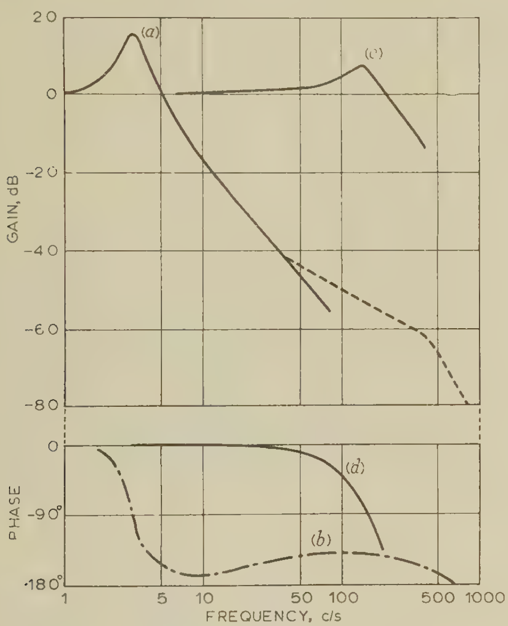


Fig. 4.—Frequency/response characteristics.

(*a*) Transducer response $|v/i_3|$.
— — — Response of phase-advancer and phase-sensitive rectifier.
(*b*) Theoretical open-loop phase characteristic.
(*c*) Closed-loop response of multiplier $|i_3/i_1|$; $i_2 = i_4 = 20$ mA.
(*d*) Closed-loop phase characteristic.

The control circuit to achieve this is quite simple and consists of a double-triode a.c. amplifier, a full-wave phase-sensitive rectifier with filter, and a d.c. amplifier. The latter is a simple operational amplifier with feedback from the output current, i_3, and input and output networks to make it into a 10:1 phase-advancing circuit giving $\frac{i_3}{v}(s) = K\left(\frac{1 + T_s}{1 + 0 \cdot 1 T_s}\right)$. The bulk of the system gain is in the a.c. amplifier and hence the change in the output current, i_3, due to drift in the d.c. amplifier is small, usually much less than ±1% of the rated output of 20 mA.

The variations of input-winding resistances with temperature are, of course, another source of error when the unit is fed from a voltage source. However, if swamp resistances are included to raise the input-circuit resistance to 1 kΩ or more, the variation of resistance is less than +0·5% for a 10° C temperature rise above an ambient of 15° C.

The overall frequency-response of the system, $(i_3/i_1)(j\omega)$, for constant i_2 and i_4, is shown by curve (*c*) in Fig. 4. There it can be seen that over the range 0–50 c/s the amplitude error is within 2 dB and the phase error is within 10°. This is the region which is exploited for an accurate analogue multiplier.

(3) PERFORMANCE AS AN ANALOGUE MULTIPLIER

(3.1) Zero Frequency Accuracy

To evaluate its performance as an analogue multiplier the unit was tested with d.c. inputs i_1 and i_2 in the range −20 to +20 mA in all four quadrants and the deviations from the ideal law of $i_3 = i_1 i_2 / i_4$ were never greater than ±1·5%. In addition, the unit was arranged to square an input current by passing it through both input windings in series, and again the errors were less than ±1·5%.

(3.2) Dynamic Accuracy

Provided that one input signal is of low frequency, say 0–2 c/s, the dynamic performance of the multiplier will be substantially that indicated by curve (*c*) of Fig. 4. If, however, the multiplier has to square a sinusoidal input signal, the effective frequency range is virtually halved. This is because the input torque consists of a steady component plus a double-frequency component of equal magnitude, and in order to develop an opposing torque for the latter the control system has to operate at twice the frequency of the input signal. Thus, when used as a squaring circuit the bandwidth of the multiplier has to be restricted to 0–25 c/s for accurate results.

The unit may be used as an indicator of the mean value of the product of the input currents, $\overline{i_1 i_2}$, by putting a moving-coil meter in the output circuit which will respond to the mean value of the i_3 waveform. In a test of this kind an input current ranging from 5 to 60 c/s was squared and the error in the mean d.c. output, $\bar{i}_3$, was within ±1% of the rated output current of 20 mA.

These results were obtained with torque motors made from 12-slot synchro laminations and so the inductance parameters are influenced by the slotting. It is expected that with motors of the pattern shown in Fig. 1(*b*) improved accuracies will be possible.

(4) APPLICATIONS

In addition to its use as an analogue multiplier this device has further applications.

The simplest of these is as an isolating device. By making $i_2 = i_4$ the output current, i_3, is a replica of i_1 over the range 0–50 c/s, but, of course, the input circuit may be isolated from the output. This is useful when a number of currents in different circuits have to be monitored and displayed on a multi-channel high-speed pen recorder with a common connection to the input channels. This occurs, for example, in testing electrical process-control circuits and the d.c. machine circuits used for large automatic-control systems.

The unit may also be used as an electrodynamic ammeter, voltmeter or wattmeter by including a moving-coil meter in the output circuit which will respond to $\overline{i_1 i_2}$, as mentioned in Section 3.2. Such a scheme could be used for a scanning system of supervisory control using the d.c. component of i_3 for telemetering purposes.

Such a unit could also be used as a power-sensitive device to give improved frequency control of the alternators in central power stations by anticipating the change in shaft speed which initiates governor action.

It may also be used as a ratio meter giving process efficiency. By keeping i_2 instead of i_4 constant and using i_1 and i_4 as the input currents, the output current, i_3, is proportional to i_1/i_4. It is, of course, essential that i_4 is greater than zero and that i_3 does not exceed the rated value for the unit.

(5) CONCLUSIONS

The multiplier, of comparatively crude construction, has given surprisingly good performance, both in terms of accuracy and frequency response, and with torque motors of the type shown in Fig. 1(*b*) an improvement in accuracy is expected.

The control circuit is extremely simple, and if transistorized the complete unit would occupy a volume no larger than $1\frac{1}{2}$ in × 3 in × $3\frac{1}{2}$ in. With a crossed-hinge suspension or a taut suspension in place of the present jewelled bearings, a device of this kind would be well suited for long periods of continuous operation.

(6) ACKNOWLEDGMENT

The author is indebted to Mr. T. Telford, who carried out the initial development work on the multiplier whilst a student of the Graduate course on Information Engineering at the University of Birmingham.

WORK FUNCTION MEASUREMENTS ON THE PLATINUM ALLOYS OF THE ALKALINE-EARTH METALS

By H. Batey.

(*Communication first received* 12*th January, and in revised form* 10*th April,* 1961.)

The thermionic work function has been measured for alloys of the alkaline-earth metals, barium, strontium and calcium, with platinum. The values obtained are comparable with published data on the photo-electric work function of the three pure alkaline-earth metals.

The alkaline-earth metals, barium, strontium and calcium, are all relatively volatile *in vacuo*, and measurements of their respective work functions cannot therefore be made by the direct thermionic method of observing the slope of a Richardson line. The usual technique employed is the photo-electric method, although a number of workers have used the contact-potential method. Details of such work have been summarized by Herrman and Wagener.[1] In a more recent survey by Wright[2] the following values of work function are quoted as representative.

	eV
Barium ..	2·5
Strontium ..	2·6
Calcium ..	3·2

It has been recognized recently that the three alkaline-earth metals alloy readily with platinum at red heat, and the object of the present note is to compare the work functions of the three binary alloys with those of the pure metals.

An alloy of platinum and barium is prepared in an evacuated system by distilling the vapour of barium from a pellet of the metal held in an electrically heated spiral of tungsten wire. The barium dissolves in the surface layers of a box-type platinum cathode core, held at 1 200° K. The thickness of the barium–platinum alloy may be several microns, as shown by micro-photographs of cross-sections of the platinum core.[3] The surface of the alloyed core has a uniform matt appearance of light-grey colour. The alloys of strontium and calcium are prepared in the same manner.

The alloyed core is removed from the vacuum system and built into a conventional diode structure, with an insulated internal tungsten heater and a close-spaced grid as electron collector or anode. The core is fitted with a thermocouple, and the whole system is sealed into the usual glass envelope fo evacuation on a high-grade bench pump. After the conventiona vacuum treatment the system is sealed from the pump and i ready for measurement.

Measurement of work function is undertaken by observin the variation with temperature of the temperature-limite current, I_s, drawn from the alloy surface. The observations ar plotted in the usual form of log I_s or log (I_s/T^2) versus 1 000/T and the work function is derived from the slope of the resultin straight line. Typical results for the three alloys are set out i Fig. 1, and details of the individual measurements on nin samples are set out in Table 1.

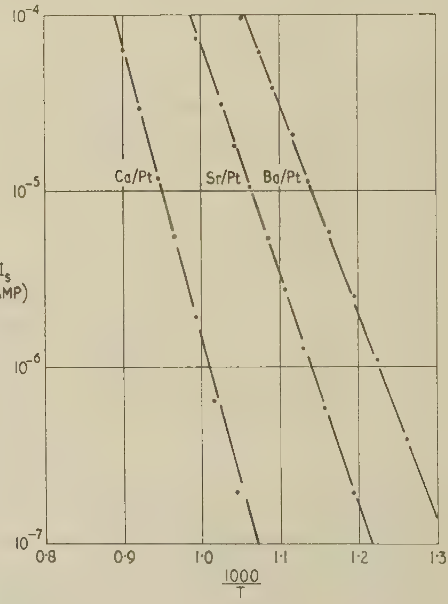


Fig. 1.—Thermionic emission from alloys of platinum with barium strontium and calcium.

Effective emitting area, 0·4 cm².

Mr. Batey is at the Post Office Research Station.

Table 1

THERMIONIC EMISSION FROM ALLOYS OF PLATINUM WITH BARIUM, STRONTIUM AND CALCIUM

Alloying element	Tube No.	Work function, ϕ	
		$I_s = A \exp(-\phi/kT)$	$I_s = AT^2 \exp(-\phi/kT)$
		eV	eV
Barium	1	2·25	2·11
	2	2·44	2·32
	3	2·30	2·13
	4	2·60	2·44
	Mean	2·4	2·2
Strontium ..	5	2·41	2·27
	6	2·56	2·36
	Mean	2·5	2·3
Calcium	7	2·70	2·53
	8	3·32	3·23
	9	3·13	3·08
	Mean	3·1	3·0

Comparison of the mean values of the work functions of the alloys with those of the alkaline-earth metals shows that they are very similar. The result suggests the possibility of direct thermionic measurement of the work functions of volatile metals when these are capable of forming relatively non-volatile alloys with platinum.

Acknowledgment is made to the Engineer-in-Chief of the Post Office for permission to make use of the information here given. The author also wishes to thank Dr. G. H. Metson for his helpful advice during the course of the work.

REFERENCES

(1) HERRMAN, G., and WAGENER, S.: 'The Oxide-Coated Cathode' (Chapman and Hall, 1951).

(2) WRIGHT, D. A.: 'A Survey of Present Knowledge of Thermionic Emitters', *Proceedings I.E.E.*, Paper No. 1404 R, November, 1952 (**100**, Part III, p. 125).

(3) METSON, G. H., and MACARTNEY, E.: 'The Conductivity of Oxide Cathodes, Part 8', *ibid.*, Monograph No. 357 E, February, 1960 (**107**, C, p. 158).

THE SUPPRESSION OF CORONA- AND PRECIPITATION-INTERFERENCE IN V.H.F. RECEPTION

By H. PAGE, M.Sc., Member.

(*Communication first received* 17*th February, and in revised form* 20*th April,* 1961.)

It has long been known that radio reception may be subject to severe interference in thundery weather conditions. Since it is usual for rain or snow to be falling at these times the interference is sometimes described as 'precipitation static', or as due to 'charged rain'. Interference of this nature is sometimes very severe on those B.B.C. circuits in which a Band I television station is picked up at another station, for rebroadcasting; as a result, the picture may frequently be seriously degraded, or quite unusable, for many minutes on end.

The following is an account (intended only as an initial progress report) of some work which has recently been started with a view to eliminating the trouble, and which has already given promising results.

TYPES OF INTERFERENCE

It is generally known that the interference occurs only when there are electrical storms in the neighbourhood, and that it manifests itself in two ways. First, charged clouds set up a high potential gradient, causing corona discharges from the receiving aerial or supporting mast. Second, rain (or other forms of precipitation) falling on the aerial may be charged. In both cases a pattern of white dots is produced on the picture, and background noise occurs on the sound programme. In severe cases the operation of the line synchronizing circuits in the receiver may be disturbed, and 'line tearing' occurs.

To investigate to what extent the interference can be reduced in practice, by relatively simple means, experiments have been carried out at the B.B.C. station at Thrumster, Caithness; the programme is derived by direct reception from Meldrum, Aberdeen, on 61·75 Mc/s, and interference is sometimes severe. Two identical receiving aerials have been set up (each comprising two tiers of horizontal Yagis) mounted at a height of 140ft on similar towers. One aerial is of conventional design and the other, which is illustrated in Fig. 1, is designed to be protected against both forms of interference by the following means:

Corona-interference.—The tower supporting the aerial is surmounted by a thin metal pole 24ft long. This is sharpened at the top to a fine point (to encourage the topmost point to discharge first), while the tips of the aerial rods are rounded to inhibit corona. The horizontal aerial rods are mounted symmetrically with respect to a plane containing the tower axis, and so theoretically do not couple with waves propagated down the tower from the discharge point. As an additional precaution, tuned rods are provided to reflect the strongest wave component (for which the electric force is radial to the mast axis), thus reducing its magnitude at the aerial. In addition, the tower was sited, and the horizontal radiation pattern of the aerial arranged, to minimize pick-up of corona-interference from other sources, particularly the main transmitting mast.

Precipitation-interference.—Each aerial rod is encased in a cylindrical insulating tube, 2in in diameter, to prevent precipitation falling directly on the rod.

The field strength of the wanted signal at the receiving aerials is 0·5mV/m, and the aerial gain is approximately 6dB relative to a half-wavelength dipole.

TEST RESULTS

The comparison between the 'protected' and 'unprotected' aerials is made by connecting them to identical monitors, and making subjective assessments of the interference. The aerials were installed in November, 1960, and during the subsequent two months there were 64 such assessments, the interference varying in duration from a few minutes to one hour. The results are summarized in the following table.

Subjective assessment of interference	Number of reports received	
	Protected aerial	Unprotected aerial
Imperceptible	49	1
Just perceptible, with careful viewing ..	12	7
Perceptible, but good entertainment value	1	20
Slightly disturbing, but fair programme value	2	10
Disturbing, poor programme value ..	0	20
Very disturbing, picture unusable ..	0	6

Mr. Page is in the Research Department, British Broadcasting Corporation.

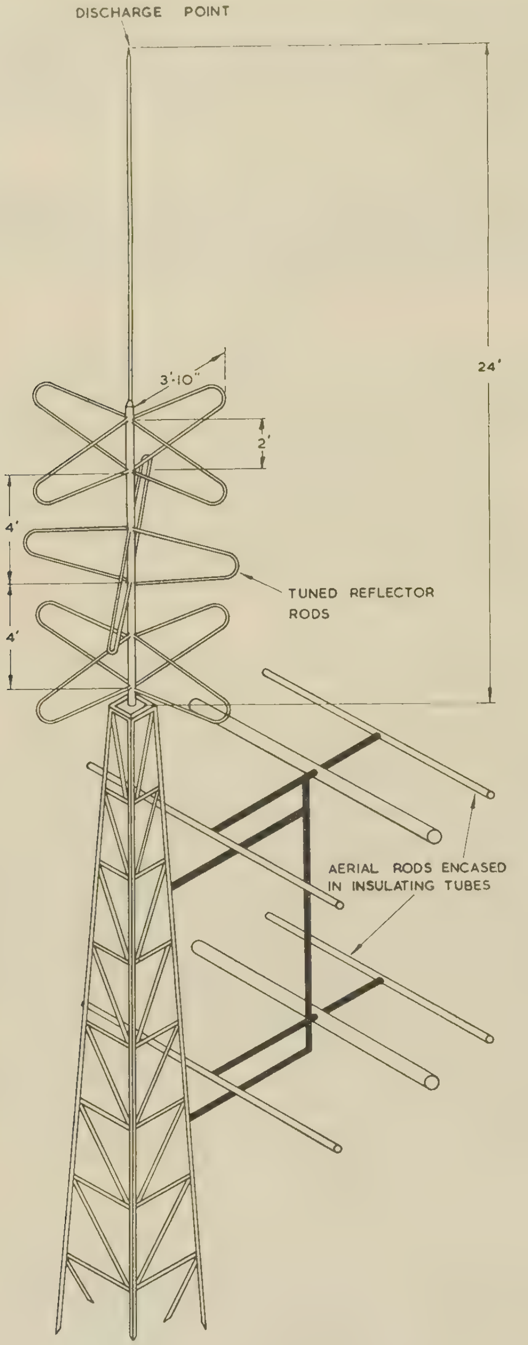
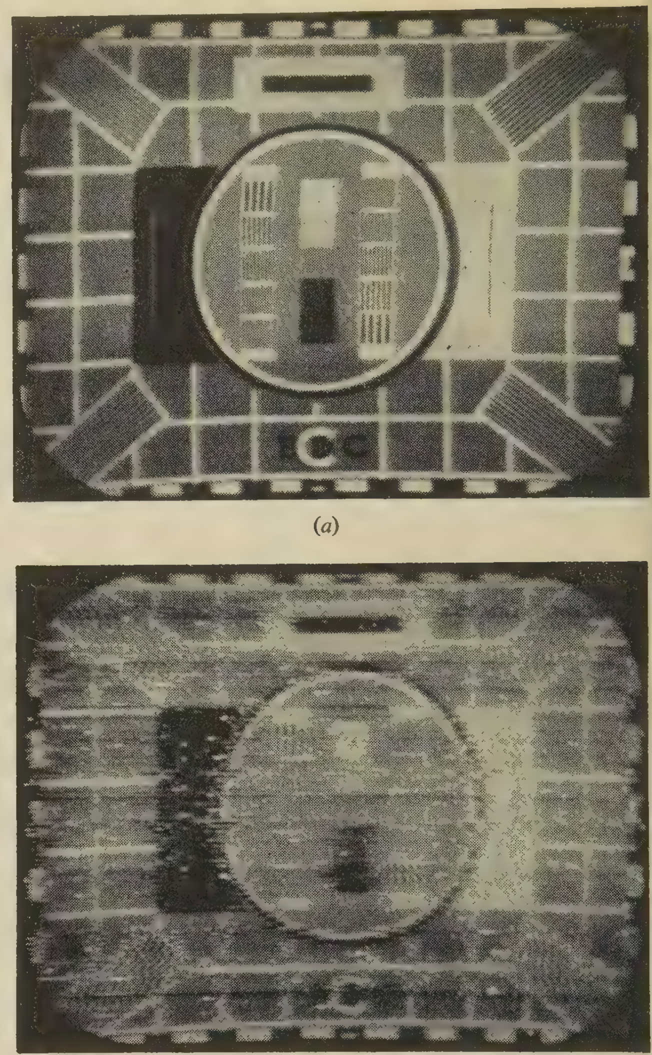


Fig. 1.—The protected aerial.

These preliminary results show that a considerable measure of success has been achieved. For example, over the period considered no instances occurred when the signal provided by the protected aerial was either unusable or of poor programme value. A comparison of the pictures on the two monitors in conditions of interference assessed as 'disturbing' is shown in Fig. 2.

A separate series of tests was made to assist in interpreting the subjective gradings, the interference level being varied to correspond to subjective assessments. It was decided that the protective measures are equivalent to a reduction of the interference level by at least 20 dB, and possibly greater in view of the large number of 'imperceptible' reports for the protected aerial.



(*a*)

(*b*)

Fig. 2.—Typical pictures when interference on unprotected aerial was assessed as 'disturbing'.

(*a*) Protected aerial.
(*b*) Unprotected aerial.

The tests are continuing with a view to obtaining more information on the nature and statistics of the phenomenon, and it is hoped in due course to publish a fuller report of the work.

The work described has been carried out by a team from the Radio Group of the B.B.C. Research Department; in particular, J. R. Chew, P. Knight, G. D. Monteath and D. J. Whythe have made valuable contributions to the investigation. Thanks are due to the operating staff at Thrumster for their co-operation in the tests, and to the B.B.C. Planning and Installation Department for providing the test aerials. The communication is published by permission of the Director of Engineering of the B.B.C.

PAPERS AND MONOGRAPHS PUBLISHED INDIVIDUALLY

Summaries are given below of papers and monographs which have been published individually. The papers are free of charge; the price of the monographs is 2s. each (post free). Applications, quoting the serial numbers as well as the authors' names, and accompanied by a remittance where appropriate, should be addressed to the Secretary. For convenience, books of five vouchers, price 10s., can be supplied.

Diversity Reception and Automatic Phase Correction. Paper No. 3584 E.

L. LEWIN.

The interference between the direct ray and a ground-reflected ray gives rise at the receiver of a communication link to a sinusoidal field pattern in the vertical plane consisting of nodes and maxima. The position and pattern wavelength of this field depend on the receiver and transmitter heights and spacing and on propagation conditions via an effective curvature parameter, C. This parameter varies in time, and is the cause of most fading at a single antenna. Its upper limit, which may be as high as $2 \cdot 5$ for a small percentage of the time, determines minimum antenna heights for line-of-sight transmission under extreme conditions. C is normally about $\frac{3}{4}$. Its range of variation determines the optimum spacing of a pair of diversity antennae, and suitable design formulae are given. Experiments using a pair of mirrors, a varying transmitter frequency and photographs of oscillograph traces indicate an extreme lower value of C over water of $-1 \cdot 5$. A moving-film display shows that conditions can vary rapidly from minute to minute, although at other times the display is steady for hours at a time.

An automatic phasing junction has been designed for insertion in the feed from two diversity antennae, the drive for the phasing element being taken from the receiver. A combined signal nowhere smaller than the greater of the received signals is obtained from the combining unit, and when the diversity spacing is chosen with regard to the extreme values of curvature obtained on the path, an excellent overall response is ensured. Some preliminary figures are given for the performance of an improved combining network and phase-control apparatus operating in conjunction with a height-diversity microwave system over a 69-mile over-water path.

Tests of Three Systems of Bandwidth Compression of Television Signals. Paper No. 3613 E.

G. F. NEWELL and W. K. E. GEDDES, M.A.

The paper describes an investigation into three similar methods of reducing the bandwidth required for the transmission of television pictures. The three methods all involve the isolation of the essential brightness changes in the picture; these are redistributed to occur at a uniform rate for transmission. A second signal must be transmitted in order to allow correct repositioning of the brightness changes in the final picture. Partial instrumentation has been carried out, and the methods have been assessed with regard to their bandwidth requirements and resultant picture quality.

Wide-Band Coupling Systems between a Waveguide and a Transmission Line. Monograph No. 442 E.

B. ROGAL, B.Sc.(Eng.), and Prof. A. L. CULLEN, O.B.E., Ph.D.

The paper describes a wide-band microwave system suitable for coupling power between a coaxial line and a waveguide or between two waveguides. Such a system can be applied to normal matched transmission as well as, for example, to coupling power between a coaxial klystron cavity and a waveguide. An appreciable improvement in performance over other schemes results in an almost constant power transfer over a bandwidth in the region of 25%. A practical example of the system in X-band is described.

The Conductivity of Oxide Cathodes. Part 10.—Spontaneous Generation of Negative Ions. Monograph No. 443 E.

G. H. METSON, M.C., D.Sc., Ph.D., M.Sc., B.Sc.(Eng.).

The experimental work described here proves that a barium-strontium-oxide matrix is thermally unstable at 1 020° K and continuously generates negative ions of oxygen in its hollow pore system. The barium-oxide component of the matrix is essentially stable and the strontium-oxide component is the oxygen-ion generator. The action is a fundamental one and proceeds at constant and unalterable rate. Factors determining the equilibrium concentration of free oxygen ions in the pores are examined and described. One conclusion reached is that donor concentration of a thermally-activated barium-strontium-oxide matrix is almost wholly strontium metal.

The Conductivity of Oxide Cathodes. Part 11.—Thermal Stability of the Alkaline-Earth Oxides. Monograph No. 444 E.

G. H. METSON, M.C., D.Sc., Ph.D., M.Sc., B.Sc.(Eng.), and H. BATEY.

The three common alkaline-earth oxides are examined for thermal stability at 1 200° C. In an experimental arrangement involving a massive evaporation from a platinum substrate it is shown that barium oxide is removed in the form of unchanged molecules while strontium and calcium oxides leave the substrate in elemental form as metal and oxygen. It is concluded that the dissociation is a thermal one and not concerned with the nature of the substrate.

Transitional Electrical Units. Monograph No. 445.

LEO YOUNG, Dr.Eng., M.A., M.S.

The paper is concerned with the six metric systems of electrical units in common use. The equations of each system differ by constants involving 4π and (approximately) 3×10^{10}. Although this is well understood and accepted, difficulties sometimes arise when a comparison is made between the units or quantities of different systems. The paper endeavours to show that these difficulties are largely semantic, and that they can be overcome with 'transitional electrical units' by converting a given unit into its transitional counterpart, and thence into the unit of the desired system.

The Launching of Surface Waves by an End-Fire Array of Slots. Monograph No. 450 E.

Prof. A. L. CULLEN, O.B.E., Ph.D., and J. A. STANIFORTH, Ph.D.

An end-fire array of slots suitable for launching a surface on a dielectric-coated metal sheet is described and analysed. The analysis is based on the representation of the slots as magnetic dipoles in the plane of the sheet, their axes being perpendicular to the line of the array.

In a practical embodiment of the scheme the elements take the form of slots fed by quarter-wavelength branch guides series fed from a waveguide partially filled with dielectric mounted beneath the metal sheet. Experimental results are given which support the theoretical conclusion that a high launching efficiency is possible. A launching efficiency of 95% is obtained theoretically for a 12-slot array launching a surface wave on a metal sheet coated with dielectric of relative permittivity $2 \cdot 56$ and of thickness $0 \cdot 125$ in at a frequency of $1 \cdot 38$ Gc/s.

Backward Waves in Waveguides containing Dielectric. Monograph No. 451 E.

P. J. B. CLARRICOATS, B.Sc.(Eng.), Ph.D.

A method is described for determining the conditions which ensure backward-wave propagation in dielectric-loaded inhomogeneous waveguide structures. It is established that backward-wave propagation can be ensured for the hybrid H_{11}-mode in a circular waveguide containing an axial dielectric rod whose relative permittivity exceeds approximately $9 \cdot 4$. The possibility of other inhomogeneous-waveguide modes exhibiting backward-wave properties is also examined.

The Relation between Discrete Periodic Inputs, the Transfer Function and the Transient Response of a System. Monograph No. 452 M.

T. GLUCHAROFF, M.E.

Sampled-data systems are characterized by the presence of discrete signals at some point of the system, but the overall output is usually a continuous function of time. It is shown that the pulse sequence of discrete periodic signals, which result in a finite-settling-time response when applied to the input of a system, can be determined directly from the system transient response. Such a pulse sequence can be used to

design a discrete controller to compensate the system, when its transfer function is not known. Further, it is shown that the transfer function of a system can be found once an input pulse sequence has been determined, with an accuracy limited only by that of the given transient response.

Passive Bistatic Radar Enhancement Devices. Monograph No. 455 E.

L. Peters, B.E.E., M.S., Ph.F.

The usual enhancement devices are constructed to give a maximum monostatic echo area. This is achieved by setting up a uniform phase distribution in the effective aperture of the device independent of the angle of incidence. Bistatic echo-area enhancement may be achieved by altering this phase distribution. Gable phase distribution is considered for the rectangular aperture and is applied to the case of the triple-bounce corner reflector. Both linear radial and quadratic radial phase distributions are treated for circular apertures and applied to the Luneberg reflector and those spherical devices described by Kay.

It is shown that the echo pattern may be approximated as an exponentially decaying function of the bistatic angle θ, i.e. $\varepsilon^{-\gamma\theta}$. In each case the parameter γ is related to some physical parameter of the device by a design process. The validity of this design process has been verified for a triangular triple-bounce corner reflector and a Luneberg reflector.

The Launching of an Axial Cylindrical Surface Wave. Monograph No. 457 E.

J. Brown, D.Sc.(Eng.), and H. S. Stachera, B.Sc.(Eng.), Ph.D.

The most usual form of launcher of an axial surface wave is a horn which acts as a mode transducer between a coaxial line and a surface waveguide.

A full mathematical treatment of the horn presents considerable difficulty and therefore, as a step towards obtaining a solution for such a mode transducer, a launcher in the form of an annular slot in a conducting plate has been investigated. An exact mathematical solution has been obtained for the electromagnetic field at a large distance from the slot. Expressions for the surface-wave field and the radiated field were obtained and used to calculate the power carried by the two waves. The efficiency of launching of a surface wave was evaluated and the results were confirmed by experimental methods.

A Critical Discussion of the Input Impedance of Idealized Mathematical Antenna Models. Monograph No. 458 E.

E. V. Bohn, Dr.rer.Nat.

A difficulty with present mathematical antenna models is the fact that there is no theoretical justification for equating the computed input impedance with those measured experimentally. It is shown that a rigorous definition of antenna impedance can be based only on a computation of the complex power in the electromagnetic field from energy integrals. To account for this power it is essential to consider the transmission line feed to the antenna, and an antenna model is derived on this basis. This model is then compared with presently accepted models. The usefulness of the induced-e.m.f. method is discussed. It is then shown that the slice generator used in the Hallen–King model does not satisfy the requirement for the conservation of power and that the impedance determined from this model depends on a semi-empirical choice of expansion parameter.

PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Part B. ELECTRONIC AND COMMUNICATION ENGINEERING (INCLUDING RADIO ENGINEERING), JULY 1961

CONTENTS

Papers for the Proceedings.—An author who supplies an outline of a paper he proposes to submit for the *Proceedings* may apply to the Secretary for a free copy of The Institution's Handbook for Authors. This gives particulars of a number of requirements—including maximum acceptable length—compliance with which is essential. See page xxii in the advertisement section.

Bibliographical References.—It is requested that bibliographical reference to an Institution paper should always include the serial number of the paper and the month and year of publication, which will be found at the top right-hand corner of the first page of the paper. This information should precede the reference to the Volume and Part.

Example.—SMITH, J.: Reflections from the Ionosphere', *Proceedings* I.E.E., Paper No. 5001 R, December, 1960 (102 B, p. 1234).

Published by The Institution, Savoy Place, London, W.C.2. *Telephone:* COVent Garden 1871. *Telegrams:* 'Voltampere, Phone, London.'
Printed by Unwin Brothers Limited, Woking and London.